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 India [2]Indian Institute of Technology Patna
Finding Needles in Images - Blog Post
{parth.thakkar, ankush.agarwal}@fujitsu.com

Abstract

While Multi-modal Large Language Models (MLLMs) have shown impressive capabilities in document understanding tasks, their ability to locate and reason about fine-grained details within complex documents remains understudied. Consider searching a restaurant menu for a specific nutritional detail or identifying a disclaimer in a lengthy newspaper article — tasks that demand careful attention to small but significant details within a broader narrative, akin to Finding Needles in Images (NiM). To address this gap, we introduce NiM-Benchmark, a carefully curated benchmark spanning diverse real-world documents including newspapers, menus, and lecture images, specifically designed to evaluate MLLMs' capability in these intricate tasks. Building on this, we further propose Spot-IT, a simple yet effective approach that enhances MLLMs capability through intelligent patch selection and Gaussian attention, motivated from how humans zoom and focus when searching documents. Our extensive experiments reveal both the capabilities and limitations of current MLLMs in handling fine-grained document understanding tasks, while demonstrating the effectiveness of our approach. Spot-IT achieves significant improvements over baseline methods, particularly in scenarios requiring precise detail extraction from complex layouts.

1 Introduction

Recent breakthroughs in Multi-modal Large Language Models (MLLMs) (Team et al., 2023; Driess et al., 2023; Peng et al., 2023; OpenAI, 2023) have fundamentally transformed how machines understand and reason about visual information. These models demonstrate remarkable capabilities in visual dialogue, scene comprehension, and answering nuanced questions about visual content. For the task of Document Visual Question Answering (DocVQA) (Mathew et al., 2021), MLLMs have emerged as particularly powerful tools, interpreting visually rich documents in ways that transcend traditional text extraction methods (Fenniak and Contributors, 2022; pdfminer, 2019), enabling question answering (QA) even in documents with complex layouts and mixed text-visual elements.

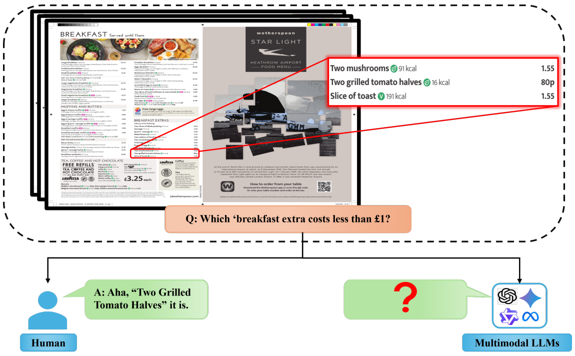


Figure 1: An example of a "Needle in Images" task: finding a specific breakfast extra under £1 in a restaurant menu requires precise attention to a small region while processing the entire layout. **How do MLLMs compare to humans on such tasks? We present a benchmark and a baseline method to study this.**

While MLLMs excel at broad document comprehension, their ability to handle precise, localized information within complex documents remains an open question. Consider a seemingly simple task: *Searching a Restaurant Menu to find a breakfast extra that costs less than £1* (as shown in Figure 1). This information occupies just a tiny fraction of the document's spatial extent, yet humans can efficiently locate it by combining broad visual scanning with focused attention – quickly zeroing in on "Two Grilled Tomato Halves" as the answer. This everyday scenario highlights a fundamental challenge in document understanding: the ability to locate and reason about minute details within larger document.

Traditional approaches based on OCR and text-extraction (Smith, 2007; Memon et al., 2020;

[*]Equal contribution as co-first authors.
[†]Work done during internship at Fujitsu Research India
Code: GitHub Repo, Data: Hugging Face Dataset

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23626–23648
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

pdfminer, 2019) inherently struggle with this challenge, as they often lose the crucial connection between local details and global document structure. Even for MLLMs, despite their broad training on web-scale data (Gadre et al., 2024), processing fine-grained details within visually rich documents presents a unique challenge, especially in domain-specific documents with complex visual layout (shown in Figure 3). This difficulty stems from a fundamental tension: models must simultaneously maintain document-level context while precisely attending to minute details – a capability that humans possess naturally but remains elusive for automated systems.

The current landscape of DocVQA research has not adequately addressed this challenge. While pioneering work like DocVQA (Mathew et al., 2021) established foundations for document understanding using MLLMs, it primarily focuses on general comprehension tasks in industrial documents. Subsequent benchmarks such as SlideVQA (Tanaka et al., 2023) and MMLongBench (Ma et al.) have expanded the scope to multi-page scenarios and long-form documents, respectively. However, these benchmarks evaluate broad document comprehension rather than the specific challenge of locating and reasoning about minute details within complex layouts. This gap is particularly significant as real-world document interaction often depends on precisely locating and interpreting small but critical pieces of information within a larger context.

To address this gap, we introduce the Needles in Images Benchmark, NiM-Benchmark. This carefully curated benchmark specifically evaluates fine-grained visual reasoning in DocVQA across diverse real-world scenarios - from dense newspaper layouts to intricate restaurant menus, magazine spreads, and classroom lecture snapshots. Each document type presents unique challenges in locating and reasoning about minute details within complex visual contexts. The benchmark includes targeted question types that probe a model's capability to combine broad document understanding with precise attention to relevant local details, closely mirroring real-world information seeking scenarios.

To complement our benchmark, we propose Spot-IT, a simple yet effective approach that draws inspiration from human visual search behavior. Our method enhances MLLMs' ability to focus on specific document regions through a novel question-guided attention mechanism. For each input document, Spot-IT segments the image into patches, identifies the most relevant regions based on the query, and dynamically generates a Gaussian patch with a variable σ, adjusted using cosine similarity (as illustrated in Figure 2). This approach enables models to better handle the dual challenges of maintaining global context while attending to local details. Below, we summarize the key contributions of our work:

1. We formalize the *Needle in an Image* challenge in DocVQA, focusing on evaluating MLLMs' ability to locate and reason about fine-grained details within complex documents.
2. We introduce NiM-Benchmark, a carefully curated benchmark comprising $2,970$ images and $1,180$ question-answer pairs across diverse document types including academic papers, newspapers, menu and images from classroom lectures. Each question is specifically designed to test MLLMs' capability to extract precise details within rich visual contexts, with rigorous quality validation through both human experts and automated verification.
3. We propose Spot-IT, a simple yet effective approach that enhances MLLMs' fine-grained reasoning capabilities through question-guided dynamic attention. Our method achieves this without requiring architectural changes to existing MLLMs, making it broadly applicable across different model architectures.
4. Through comprehensive experiments, we demonstrate that Spot-IT significantly improves state-of-the-art on fine-grained detail extraction, achieving a 15.5% improvement over GPT-4o on ArxiVQA and 21.05% improvement on our NiM-Benchmark. These results establish new baselines for precise information extraction in DocVQA .

2 Background and Related Work

Document Understanding Evolution: Document understanding has evolved from rule-based OCR systems (Smith, 2007; Subramani et al., 2020) to sophisticated Multi-modal Large Language Models (Team et al., 2023; OpenAI, 2023). Early DocVQA datasets (Mathew et al., 2021; Du et al., 2022) focused on basic text extraction and comprehension tasks, while recent benchmarks like SlideVQA (Tanaka et al., 2023) and MMLongBench (Ma et al.) have expanded to multi-page scenarios and long-form documents. However, these datasets primarily evaluate broad document comprehension rather than fine-grained detail extraction, which is the primary motivation for creating our benchmark. We compare our benchmark with existing ones in Table 3 (in Appendix).

Fine-grained Visual Analysis in Documents: While fine-grained visual analysis has been extensively studied in natural images (Yang et al., 2023), its application to document understanding remains limited. Recent visual prompting techniques (Wu et al., 2024) have shown promise in directing model attention to specific image regions through bounding boxes (Lin et al., 2024) or markers (Shtedritski et al., 2023). However, documents present unique challenges due to their hierarchical structure and complex layouts, making direct adaptation of these techniques insufficient. Our work bridges this gap by introducing both a benchmark and method specifically designed for evaluating fine-grained document analysis capabilities of MLLMs.

Methods for Document VQA: Current approaches to DocVQA either rely on traditional OCR-based pipelines (Xu et al., 2020b; Huang et al., 2022) or leverage end-to-end MLLMs (Zhang et al., 2024b,a). For larger documents, retrieval-augmented generation (RAG) methods (Faysse et al., 2024b) have emerged as a promising direction. However, these methods typically process entire document regions without considering the granularity of relevant information, leading to inefficiencies when only small portions contain the answer. Our Spot-IT addresses this limitation through a question-guided attention mechanism that selectively focuses on relevant document regions. For an extended discussion of related work, please refer to Appendix A.1.

3 Dataset: Needle in an Image Benchmark

Our benchmark, NiM-Benchmark, is designed to evaluate MLLMs' ability to locate and reason about fine-grained details within complex documents. We define fine-granularity using the following ratio:

$$\text{Fine-Granularity} = \frac{\text{Area of Relevant Region}}{\text{Total Image Page Area}} < 0.05 \quad (1)$$

i.e., a task is considered fine-grained when the relevant region occupies less than 5% of the total image area.

In this section, we describe the dataset construction process, its characteristics, and provide an in-depth analysis.

3.1 Dataset Construction

Our dataset spans multiple domains including academic papers, newspapers, magazines, lecture materials, and restaurant menus, each presenting unique challenges in locating fine-grained information.

Document Collection and Processing: We curated documents from six diverse domains: (1) Restaurant menus with complex layouts and pricing information, (2) Recent academic papers from arXiv (2024-2025), (3) Magazines covering diverse domains with mixed text-visual content, (4) Contemporary English e-newspapers, (5) Website screenshots from the CoVA dataset (Kumar et al., 2022), and (6) Classroom lecture screenshots from open educational resources. Details of the domain sources are present in Table 8 (in Appendix).

To ensure consistency, all documents were converted to a uniform image format while preserving visual complexity and layout using a Python library (Belval, 2024). Documents example images are shown in Table 11 in Appendix.

Question-Answer Pair Generation: We employed a hybrid approach to create high-quality question-answer pairs that specifically target fine-grained information: (1) We divided each document into variable-sized patches (2×2 to 6×6 grids) and used a MLLM with carefully crafted prompts to generate initial QA pairs focusing on localized information within each patch (2) The initial pool of QA pairs are verified by a human annotator and the irrelevant pairs were discarded. For certain domains, automated generation with filtering proved insufficient, so a team of four annotators created fine-grained questions for those domains. (3) All QA pairs underwent verification by three independent annotators to ensure accuracy, relevance, and consistency with our focus on fine-grained detail extraction. All prompts used for dataset construction are detailed in Section A.9 in the Appendix.

3.2 Dataset Characteristics and Analysis

Our dataset includes 284 documents across six domains, containing 1,180 question-answer pairs. An overview is provided in Table 7. Each domain presents unique challenges for fine-grained information extraction, from dense multi-column newspaper layouts to technical diagrams in academic papers.

Question Types and Distribution: We categorize questions into several types to assess fine-grained understanding: (1) *Inline*: Direct extraction of specific details, (2) *Boolean*: Yes/no questions about specific details, (3) *Comparative*: Comparison between nearby elements, (4) *Complex Reasoning*: Multi-step inference about document details, (5) *Commonsense*: Requiring world knowledge, and (6) *Unanswerable*: Context needed to answer is absent. Table 9 in Appendix presents the distribution of question categories across domains.

3.3 Quality Analysis

To validate the quality of our automatically generated question-answer pairs, we conducted rigorous evaluations using two carefully curated test sets: (1) Set X containing 200 human-generated questions from existing datasets, and (2) Set Y comprising 200 samples from our dataset with balanced representation across domains (30-35 questions per domain). Our analysis encompassed three complementary dimensions:

Response Time Analysis: We measured response times and accuracy (EM and F1 scores) across three MLLMs (GPT-4o, Gemini-1.5-Flash, GPT-4o-mini) and human experts on Set Y. This analysis, visualized in Figure 5, demonstrates that although human accuracy is moderately high on our dataset, it comes at the cost of increased response time.

Question Quality Assessment: We conducted a blind Turing test where two independent researchers evaluated a mixed set of human and machine-generated questions (Sets X and Y combined). The inter-annotator agreement (Cohen's k (Cohen, 1960) = 0.234) indicates that our generated questions are comparable to human-crafted ones in terms of quality and naturalness.

Automated Verification: To ensure scalable quality assessment, we employed Claude-3.5-Sonnet and Gemini-2.0-Flash as independent judges, achieving strong inter-model agreement (k = 0.339). These models were specifically chosen to avoid potential biases, as they were not involved in the question generation process.

4 Methodology: Spot-IT

Finding a "needle" of information in a complex document requires a delicate balance between broad context awareness and precise attention to detail. Our method, Spot-IT, draws inspiration from how humans efficiently locate specific details in documents: first identifying potentially relevant regions based on the query, then focusing attention on those regions while maintaining awareness of the surrounding context. This two-stage approach enables effective extraction of fine-grained information while preserving the document's structural context.

At its core, the goal of Spot-IT is to make MLLMs focus on specific document regions through a query-guided attention mechanism. Given a document image and a query seeking fine-grained information, our method first divides the image into a grid of patches and identifies the most relevant patch using semantic similarity between the query and visual content. It then generates an adaptive Gaussian attention mask centered on this region, effectively highlighting the "needle" while maintaining visibility of the surrounding context. This attended image, along with the original query, is then processed by an MLLM to generate the final answer. Figure 2 illustrates this process.

4.1 Problem Formulation

The task of finding fine-grained details in documents can be formalized in both closed-domain and open-domain settings. In the closed-domain setting, given a query q and a document D containing a set of page images $\{I_1, ..., I_j\}$, the goal is to locate the specific region within these images that contains the answer to q. The open-domain setting extends this to a collection of documents $S = \{D_1, ..., D_M\}$, where we must first identify the relevant documents and pages before locating the specific region. In the open-domain setting, the top-r relevant documents are retrieved using methods such as ColPali (Faysse et al., 2024a), and then passed to the MLLM L.

Formally, our objective is to learn a function f that maps a query q and a set of k images $\{I_1, I_2, \ldots, I_k\}$ to corresponding attention masks $\{M_1, M_2, \ldots, M_k\}$ that highlight the regions most likely to contain the answer:

$$\{M_1, M_2, \ldots, M_k\} = f(q, \{I_1, I_2, \ldots, I_k\}) \quad (2)$$

The attended images $\{I_{M_1}, I_{M_2}, \ldots, I_{M_k}\}$ are then provided to an MLLM L along with the query to generate the answer:

$$\text{answer} = L\left(q, \{I_{M_1}, I_{M_2}, \ldots, I_{M_k}\}\right) \quad (3)$$

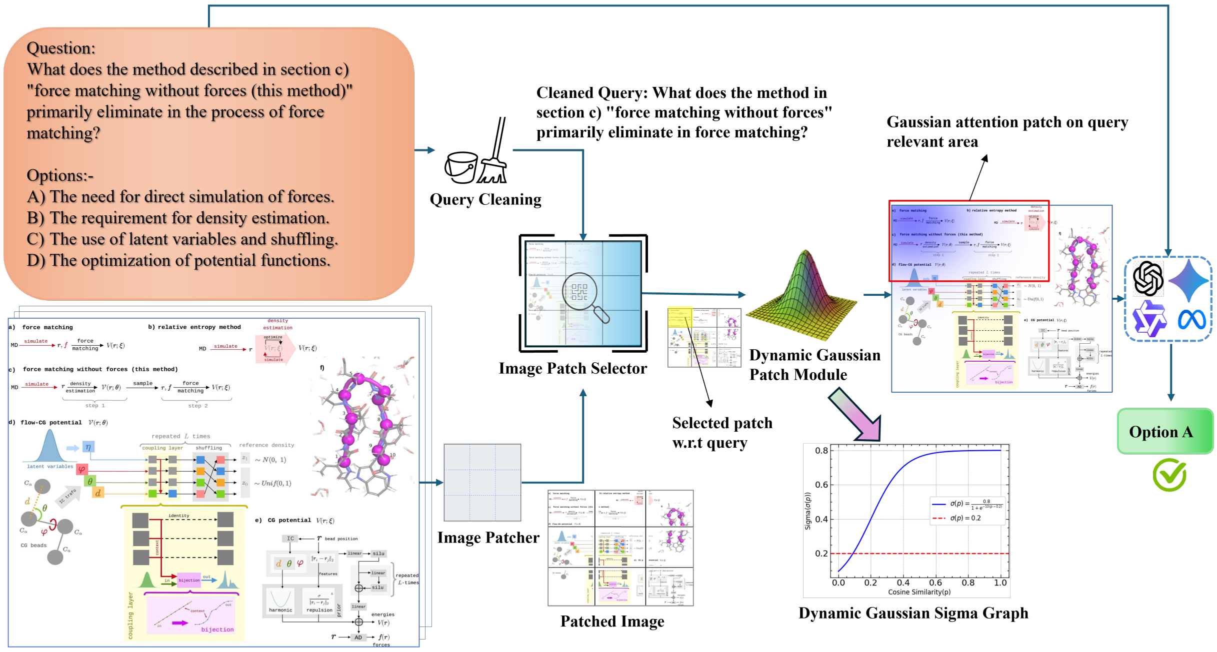


Figure 2: Overview of Spot-IT: Given a document and query, our method (1) cleans the query, (2) identifies the most relevant image patch, (3) applies an adaptive Gaussian attention mask, and (4) provides the attended image to an MLLM for answer generation. **Our method combines targeted patch selection with dynamic attention to mimic human-like focus on relevant document regions.**

The key challenge lies in designing f to effectively identify and highlight small regions containing critical information while maintaining sufficient context for the MLLM to reason about the answer.

4.2 Method Overview

Spot-IT addresses the challenge of fine-grained detail extraction through a modular pipeline that mimics human visual search behavior. As illustrated in Figure 2, our method consists of two key components:

Query-Guided Patch Identification: First, we divide the input document image into an $n \times n$ grid of patches. Using a vision-language model (SigLip (Zhai et al., 2023)), we compute semantic similarity between the query and each patch to identify the region most likely to contain the answer. This step is analogous to how humans quickly scan a document to locate relevant sections based on visual and semantic cues.

Adaptive Gaussian Attention: Once the most relevant patch is identified, we generate a Gaussian attention mask centered on this region. The spread of this Gaussian distribution adapts dynamically based on the confidence of our patch selection - higher confidence leads to more focused attention, while lower confidence results in broader attention. This mechanism directs the MLLM's focus to the identified region while preserving awareness of the surrounding context, similar to human attention.

The final attended image, created by applying this adaptive Gaussian mask to the original document, serves as input to an MLLM along with the original query. This approach enables the model to efficiently process fine-grained details within the highlighted region while maintaining awareness of the document's overall context, leading to more accurate answers for queries about specific details.

4.3 Query-Guided Patch Identification

The first key challenge in locating fine-grained information is identifying which region of the document to focus on. Our patch identification approach combines grid-based image segmentation with semantic similarity matching to efficiently locate regions relevant to the query.

Image Segmentation: Given an input document image I of dimensions $W \times H$, we divide it into an $n \times n$ grid of uniform patches. Each patch P_{ij} $(i, j \in \{1, ..., n\})$ represents a distinct region of the document. Through empirical analysis on our benchmark dataset, we found that $n = 6$ provides an effective balance between granularity and computational efficiency.

Query-Patch Similarity: To identify the most relevant patch, we leverage the SigLip vision-language model to compute semantic similarity between the query and each patch. First, we preprocess the query q by removing stop words and extraneous information to obtain a cleaned query q_c, focusing on key semantic elements. The SigLip model then encodes both the cleaned query and each patch into embedding vectors:

$$v_q = \text{SigLip}(q_c), \quad v_{ij} = \text{SigLip}(P_{ij}) \quad (4)$$

The relevance of each patch to the query is determined by computing the cosine similarity between their respective embeddings:

$$\text{Sim}(v_{ij}, v_q) = \frac{v_{ij} \cdot v_q}{\|v_{ij}\|\|v_q\|} \quad (5)$$

Patch Selection: The patch with the highest similarity score is selected as the center for our attention mechanism:

$$(i^*, j^*) = \arg\max_{i,j} \text{Sim}(v_{ij}, v_q) \quad (6)$$

To normalize the similarity score of the selected patch, we apply a softmax function over all similarity scores and define the probability p of the selected patch as:

$$p = \frac{\exp(\text{Sim}(v_{i^*j^*}, v_q))}{\sum_{i,j} \exp(\text{Sim}(v_{ij}, v_q))} \quad (7)$$

The center coordinates (x^*, y^*) of this patch in the original image space are computed as:

$$x^* = \frac{(2j^* - 1)W}{2n}, \quad y^* = \frac{(2i^* - 1)H}{2n} \quad (8)$$

This patch identification process effectively narrows down the region of interest while maintaining computational efficiency. The similarity score of the selected patch also serves as a confidence measure that influences the subsequent attention mechanism, allowing our method to adapt its focus based on the strength of the match between query and content.

4.4 Adaptive Gaussian Attention

Once we identify the most relevant patch, the next challenge is to create an attention mechanism that effectively highlights this region while preserving contextual information. We achieve this through an adaptive Gaussian attention mask that automatically adjusts its focus based on the confidence of our patch selection.

Dynamic Gaussian Mask: We generate a Gaussian attention mask centered at the coordinates (x^*, y^*) identified in the previous step. The spread of this Gaussian distribution is controlled by its standard deviation σ, which we compute adaptively based on the similarity score p of the selected patch:

$$\sigma = \frac{0.8}{1 + \exp(-10(p - 0.2))} \quad (9)$$

This sigmoid-based formulation ensures that σ varies smoothly with our confidence in the patch selection: high similarity scores result in a broader attention mask (large σ), reflecting our confidence in finding the answer in that region. In contrast, lower similarity scores yield a smaller mask, as we are less certain about the answer's location, and if σ falls below 0.2, we omit drawing a patch altogether. The parameters of this function were determined through empirical analysis on validation set of our benchmark dataset and existing datasets (see "Dynamic Gaussian Sigma Graph" in Figure 2).

Attention Mask Generation: The Gaussian attention mask $M(x, y)$ (Wu et al., 2019) for each pixel coordinate (x, y) in the image is computed as:

$$M(x, y) = \exp\left(-\frac{(x - x^*)^2 + (y - y^*)^2}{2\sigma^2}\right)^{0.5} \quad (10)$$

The square root operation in the exponent helps create a more gradual falloff in attention, which we found empirically to work better with MLLMs' visual processing capabilities.

Image Enhancement: The final attended image I' is created by blending the original image with a highlight color using the attention mask:

$$\begin{aligned} I'(x, y) =& (1 - \alpha M(x, y))I(x, y) \\ &+ \alpha M(x, y)H(x, y) \end{aligned} \quad (11)$$

where α is a blending factor (set to 0.5 in our experiments) and $H(x, y)$ represents the highlight color. This approach ensures the highlighted region remains readable and distinct.

The resulting attended image preserves the document's full content while drawing the MLLM's attention to the region most likely to contain the answer. This balance between focused attention and context preservation is crucial for accurately answering questions about fine-grained details in complex documents.

5 Spot-IT: Experimental Setup

5.1 Experimental Datasets

Existing DocVQA Datasets We evaluate Spot-IT on two DocVQA datasets: ArxiVQA (Li et al., 2024a) and DUDE (Van Landeghem et al., 2023). For evaluation, we use questions, context images, and gold answers from the ArxiVQA training set (since only the training set is available) and the

DUDE development set. Hyperparameters are tuned by randomly selecting 50 questions from each dataset. Our test set includes 500 questions from ArxiVQA and 500 from DUDE.

NiM-Benchmark For the evaluation on NiM-Benchmark, we select 937 samples distributed across the following domains: Newspapers (174), Menus (180), Lecture Screenshots (70), Website Screenshots (215), Academic Papers (180), and Magazines (118).

5.2 Spot-IT Baselines

Our approach operates in a training-free, zero-shot setting. We evaluate it against two baseline methods: an Optical Character Recognition (OCR)-based pipeline (Mishra et al., 2019) and the MLLM-DocVQA approach (Cho et al., 2024a). To ensure a comprehensive evaluation, we utilize three closed-source MLLMs—GPT-4o (OpenAI et al., 2024), GPT-4o-mini (OpenAI et al., 2024), and Gemini-1.5-flash (Team et al., 2024)—and two open-source MLLMs—Qwen2-VL 7B (Wang et al., 2024) and Llama-3.2-11B-Vision (Grattafiori et al., 2024). We additionally assess performance under Chain-of-Thought (CoT) prompting (Wei et al., 2022). This diverse selection ensures a broad and representative evaluation across both open-source and closed-source models.

OCR-Based Pipeline In this pipeline, text is first extracted from a set of images using OCR (Mishra et al., 2019), adapted from MMLongBench (Ma et al.). The extracted text is then input to the LLM, along with the corresponding question, enabling the LLM to generate an answer.

MLLM-Based DocVQA This pipeline utilizes MLLMs as the VQA model, where both the question and the corresponding context images are directly input into the model to generate an answer, as adapted from Cho et al. (2024a).

5.3 Evaluation Metrics

We use Exact-Match (EM), F1-Score (Rajpurkar, 2016), and ANLS Score (Biten et al., 2019) as automatic metrics to assess the correctness of the predicted answers. For ArxiVQA, being a multiple-choice question dataset, we use accuracy as the evaluation metric.

For NiM-Benchmark, we also conduct human evaluation on 100 samples, with the assistance of three annotators.

5.4 Implementation Details

Problem Setting: We evaluate our method in both open-domain and closed-domain settings. We use DUDE as closed-domain and convert ArxiVQA to open-domain by collating the context of all instances.

Open-Domain: The top-k most relevant images are retrieved from the corpus to answer queries, using the ArxivQA dataset.

Closed-Domain: Queries are answered using a predefined set of images that contain the exact query context, evaluated on the DUDE dataset.

Distractor Setting: Our benchmark, NiM-Benchmark, introduces distractor images to assess model resilience against irrelevant information.

These diverse settings enable a comprehensive evaluation of our proposed method against baseline models.

Context Images and MMLLMs Used: We use the same set of images across both OCR and MLLM baselines—either for text extraction or as direct inputs to the language model for answering queries. Additionally, we employ same language models for both OCR-based and image-based inputs to ensure consistency and fair comparison.

Spot-IT Hyperparameters: For query cleaning, we employ the same Multi-modal Large Language Models (MLLMs) used in the DocVQA task. The image is segmented into a 6×6 grid of patches to determine the regions relevant to the query. The standard deviation σ for the 2D Gaussian spread is selected within the range $[0, 0.8]$, as values exceeding 0.8 encompass a substantial portion of the image, thereby negating the intended effect.

For visualization, patches are highlighted using Blue color, and alpha blending is applied with a blending factor of $\alpha = 0.5$. Additionally, we impose a threshold of $\sigma < 0.2$, ensuring that if the final σ falls below this threshold, no patch is drawn. This prevents visualization in cases where the model's confidence in patch relevance is insufficient.

Experiments were performed using two NVIDIA A30 GPUs (24GB each) and MLLMs inference APIs.

6 Results and Analysis

This section is divided into two parts:

(1) *Spot-IT Evaluation:* We present the results of Spot-IT using three closed-source models—GPT-4o, GPT-4o-mini, and Gemini-1.5-Flash—and

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**

Table 1: **Spot-IT evaluation** results compared with baselines adapted from M3DocRAG (Cho et al., 2024b). **Our method outperforms all baselines, including CoT (Wei et al., 2022).**

two open-source models—Llama-3.2-VL-11B and Qwen2-7B on ArxiVQA and DUDE datasets. This is followed by an occlusion sensitivity analysis and a detailed error analysis of Spot-IT.

(2) *NiM-Benchmark Evaluation:* We assess the performance of NiM-Benchmark on GPT-4o, GPT-4o-mini, Gemini-1.5-Flash, Qwen2-7B, and human evaluators. This is followed by an error analysis of the NiM-Benchmark evaluation.

6.1 Evaluation on Document Visual QA

Table 1 presents zero-shot results on ArxiVQA and DUDE, comparing our method Spot-IT to baselines. Spot-IT consistently outperforms all baselines, including OCR and CoT, highlighting its effectiveness in efficiently finding the "needle" in the set of images. We also test our method with the proposed dataset NiM-Benchmark, achieving the best performance across all domains in various MLLM models, shown in Table 2.

Additional Results We present further evaluations on additional DocVQA datasets using GPT-4o, alongside extensive ablation studies and patch count analyses, to demonstrate the robustness and generalizability of our Spot-IT framework. Our experiments show that SigLIP consistently outperforms CLIP for patch-query similarity, and that varying the number of patches reveals an optimal trade-off between performance and generalization. Spot-IT achieves consistent gains across multiple benchmarks—DocVQA, InfoVQA, ChartQA, and MMlongbench-doc—for both short and long documents. Detailed results are provided in Appendix A.2.

6.2 Our NiM-Benchmark Evaluation

Automatic Evaluation Table 2 shows the evaluation of our proposed dataset NiM-Benchmark across SoTA MLLMs using EM, F1, and ANLS. These models exhibit low performance both on the overall benchmark and across individual domains, including Restaurant Menus, Newspapers, Website Screenshots, and Lecture Screenshots. This highlights the need to enhance MLLMs and DocVQA methodologies for locating and reasoning about fine-grained details within documents.

Human Evaluation We evaluate NiM-Benchmark using human performance, achieving 63% EM and 70% F1, highlighting significant room for improvement compared to MLLMs (Figure 4 in Appendix).

6.3 Analysis of Spot-IT

For our method, we perform: a) *Occlusion Sensitivity Analysis* - to understand model behavior, b) *Error Analysis* - to interpret failure cases, and c) *Accuracy vs. Latency Trade-off Analysis* - comparing our method with baselines.

Sensitivity Analysis

Figure 6 shows the occlusion sensitivity analysis of Spot-IT on the Qwen2-VL model. By systematically occluding image regions, the analysis identifies areas most influential to the model's predictions. Details of the occlusion methodology are in Appendix A.4.

Findings: Our method effectively highlights critical image regions that contribute to the model's predictions. This is validated by the occlusion sensitivity analysis, confirming alignment between our method's attributions and the model's decision-making process.

Error Analysis

We analyze our method on ArxivQA using GPT-4o on 500 samples, of which 200 were incorrect. We randomly selected 50% of these errors and categorized them as follows: a) *Dataset Errors* - 19%, b) *Retrieval Errors* - 22%, c) *Patch Formation* - 25%, d) *Patch Selection* - 26%, and e) *MMLLM Fault* - 8%. For details, refer Section A.5 in the Appendix.

Accuracy vs Latency Trade-off

The accuracy-latency trade-off plot compares our method with the baseline using GPT-4o on (a) ArxiVQA, (b) DUDE, and (c) NiM-Benchmark, showing a 10-20% accuracy improvement across all datasets with only an additional latency of approximately 4 seconds (see Figure 4 in Appendix).

6.4 Analysis of NiM-Benchmark

For NiM-Benchmark, we conduct: a) *Error Analysis*, and b) *Human Evaluation* to compare accuracy

Methods	Menus	Academic Papers	Magazines	Newspaper	Website Screenshots	Lectures	All
Exact Match (EM) (↑)							
GPT-4o	0.33	0.41	0.55	0.28	0.42	0.26	0.38
GPT-4o-mini	0.25	0.23	0.47	0.24	0.34	0.24	0.29
Gemini-1.5-Flash	0.22	0.17	0.19	0.14	0.30	0.34	0.22
Qwen2-7B	0.12	0.11	0.05	0.06	0.01	0.11	0.07
GPT-4o + Ours	**0.47**	**0.51**	**0.64**	**0.33**	**0.46**	0.29	**0.46**
GPT-4o-mini + Ours	0.37	0.26	0.49	0.30	0.39	0.27	0.35
Gemini-1.5-Flash + Ours	0.35	0.23	0.20	0.16	0.34	**0.41**	0.27
Qwen2-7B + Ours	0.21	0.15	0.03	0.07	0.04	0.20	0.11
F1 (↑)							
GPT-4o	0.35	0.59	0.72	0.39	0.50	0.31	0.48
GPT-4o-mini	0.25	0.38	0.62	0.35	0.42	0.32	0.38
Gemini-1.5-Flash	0.22	0.29	0.25	0.20	0.36	0.40	0.28
Qwen2-7B	0.16	0.19	0.07	0.11	0.01	0.12	0.10
GPT-4o + Ours	**0.50**	**0.66**	**0.77**	**0.44**	**0.56**	0.37	**0.56**
GPT-4o-mini + Ours	0.38	0.41	0.64	0.37	0.49	0.36	0.44
Gemini-1.5-Flash + Ours	0.35	0.36	0.29	0.20	0.40	**0.47**	0.34
Qwen2-7B + Ours	0.27	0.24	0.06	0.10	0.04	0.20	0.15
ANLS (↑)							
GPT-4o	0.55	0.61	0.71	0.49	0.55	0.39	0.56
GPT-4o-mini	0.35	0.44	0.64	0.45	0.47	0.37	0.46
Gemini-1.5-Flash	0.29	0.40	0.35	0.32	0.47	0.42	0.37
Qwen2-7B	0.19	0.29	0.18	0.25	0.08	0.16	0.19
GPT-4o + Ours	**0.63**	**0.67**	**0.78**	**0.52**	**0.60**	0.45	**0.62**
GPT-4o-mini + Ours	0.49	0.46	0.67	0.46	0.51	0.40	0.50
Gemini-1.5-Flash + Ours	0.40	0.46	0.39	0.32	0.41	**0.49**	0.40
Qwen2-7B + Ours	0.26	0.32	0.17	0.23	0.11	0.23	0.22

Table 2: **NiM-Benchmark Performance** across different domains including Newspapers, Website Screenshots, and Lectures. While Spot-IT consistently outperforms baseline models, the overall performance remains modest, highlighting the challenging nature of the benchmark and the need for further research and model improvements.

and latency with model predictions.

NiM-Benchmark Error Analysis
We evaluate the performance of NiM-Benchmark on GPT-4o by randomly selecting 20 samples from all 6 domains domain and categorized them as follows: a) *Incomplete Evidence* - 47 cases, b) *Hallucinated Evidence* - 28 cases, c) *Perceptual Error* - 24 cases, d) *Reasoning Error* - 15 cases, e) *Irrelevant Answer* - 5 cases, and f) *Knowledge Lacking* - 1 case. Refer Section A.6 in Appendix for details.

Human vs Model: Accuracy & Latency
We compare human and model performance on accuracy and latency for NiM-Benchmark. While humans achieve higher accuracy, they take significantly more time than models, highlighting the need for improved methodologies to efficiently handle our dataset (see Figure 5 in Appendix).

7 Conclusion

In this paper, we formalize the *Needle in Images* challenge in DocVQA, focusing on evaluating MLLMs' ability to locate and reason about fine-grained details within complex documents. To address this, we introduce NiM-Benchmark, a benchmark specifically designed to assess MLLMs' effectiveness in extracting precise information from visually rich layouts. Our experiments reveal that current MLLMs struggle with accurately locating and extracting answers from such intricate structures. To overcome this, we propose Spot-IT, which intelligently identifies relevant regions within images, achieving substantial improvements over baseline models across multiple datasets. We believe our findings pave the way for more advanced and efficient DocVQA systems capable of fine-grained detail extraction from complex documents.

Limitations

The limitations of our work are as follows: 1) Although our method performs well on existing DocVQA datasets, it struggles with long length documents as LLMs have limitations in processing large documents even after identifying the relevant patch. 2) The performance of our method depends on the current capabilities of LLMs, which may improve over time. 3) While achieving high accuracy, our method incurs slightly higher latency due to Gaussian patch construction. 4) We use SigLip for cosine similarity between document patches and the query using a bag-of-words-like approach, which limits contextual understanding of document structure; future work could explore a customized model for better similarity assessment. 5) Our benchmark has fewer complex reasoning questions, which can be expanded in future iterations.

Acknowledgement

We thank Pranoy Panda and other members of the AI Lab at Fujitsu Research of India for their valuable feedback on the manuscript. We are also grateful to the anonymous ARR reviewers, the meta-reviewer, and the ACL program chairs for their insightful comments, which helped us improve the paper.

References

Edouard Belval. 2024. Pdf to image library. https://pypi.org/project/pdf2image/.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024a. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024b. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *Preprint*, arXiv:2411.04952.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: an embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, pages 8469–8488.

Qinyi Du, Qingqing Wang, Keqian Li, Jidong Tian, Liqiang Xiao, and Yaohui Jin. 2022. Calm: commensense knowledge augmentation for document image understanding. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3282–3290.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024a. Colpali: Efficient document retrieval with vision language models. *arXiv preprint arXiv:2407.01449*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024b. Colpali: Efficient document retrieval with vision language models. *Preprint*, arXiv:2407.01449.

Mathieu Fenniak and PyPDF2 Contributors. 2022. The pypdf2 library.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091.

Anurendra Kumar, Keval Morabia, William Wang, Kevin Chang, and Alex Schwing. 2022. CoVA: Context-aware visual attention for webpage information extraction. In *Proceedings of The Fifth Workshop on e-Commerce and NLP (ECNLP 5)*, pages 80–90, Dublin, Ireland. Association for Computational Linguistics.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *Preprint*, arXiv:2403.00231.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2024. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. *Preprint*, arXiv:2403.20271.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE access*, 8:142642–142668.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and Aditya Ramesh et al. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5).

pdfminer. 2019. pdfminer.six.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

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*.

P Rajpurkar. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. 2023. What does clip know about a red circle? visual prompt engineering for vlms. *Preprint*, arXiv:2304.06712.

Ray Smith. 2007. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pages 629–633. IEEE.

Sargur Srihari, Stephen Lam, Venu Govindaraju, Rohini Srihari, Jonathan Hull, and E Yair. 1992. Document understanding: Research directions. In *DARPA Document Understanding Workshop, Xerox PARC, Palo Alto, CA*. Citeseer.

Nishant Subramani, Alexandre Matton, Malcolm Greaves, and Adrian Lam. 2020. A survey of deep learning approaches for ocr and document understanding. *arXiv preprint arXiv:2011.13534*.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13636–13645.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, and Libin Bai et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2023. Hierarchical multimodal transformers for multipage docvqa. *Pattern Recognition*, 144:109834.

Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. 2023. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. 2019. Gp-gan: Towards realistic high-resolution image blending. *Preprint*, arXiv:1703.07195.

Junda Wu, Zhehao Zhang, Yu Xia, Xintong Li, Zhaoyang Xia, Aaron Chang, Tong Yu, Sungchul Kim, Ryan A. Rossi, Ruiyi Zhang, Subrata Mitra, Dimitris N. Metaxas, Lina Yao, Jingbo Shang, and Julian McAuley. 2024. Visual prompting in multimodal large language models: A survey. *Preprint*, arXiv:2409.15310.

Xinya Wu, Duo Zheng, Ruonan Wang, Jiashen Sun, Minzhen Hu, Fangxiang Feng, Xiaojie Wang, Huixing Jiang, and Fan Yang. 2022. A region-based document vqa. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4909–4920.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020a. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

Lingfeng Yang, Yueze Wang, Xiang Li, Xinlong Wang, and Jian Yang. 2023. Fine-grained visual prompting. *Preprint*, arXiv:2306.04356.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024a. Cfretdvqa: Coarse-to-fine retrieval and efficient tuning for document visual question answering. *arXiv preprint arXiv:2403.00816*.

Jinxu Zhang, Yongqi Yu, and Yu Zhang. 2024b. Cream: coarse-to-fine retrieval and multi-modal efficient tuning for document vqa. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 925–934.

Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. 2022. Towards complex document understanding by discrete reasoning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4857–4866.

A Appendix

In this section, we provide detailed related work and additional results and analysis that we could not include in the main paper due to space constraints. In particular, this appendix contains the following:

- Extended Related Work
- Additional results
- Additional Figures and Tables
- Occlusion Sensitivity Analysis
- Extended Spot-IT Error Analysis
- Extended NiM-Benchmark Error Analysis
- Spot-IT Qualitative Examples
- Sample Illustrations from NiM-Benchmark
- All LLM Prompts Used for Evaluation and Dataset Generation

A.1 Extended Related Work

A.1.1 Evolution of Document Visual Question Answering

Document understanding has evolved significantly from its origins in rule-based systems (Srihari et al., 1992) and traditional OCR approaches (Subramani et al., 2020). Early systems focused primarily on text extraction and basic layout analysis (Smith, 2007), with limited ability to handle complex visual elements or perform sophisticated reasoning. The field has since transformed with the advent of MLLMs (Team et al., 2023; Driess et al., 2023; Peng et al., 2023; OpenAI, 2023), which have enabled more nuanced document understanding and reasoning capabilities.

A.1.2 DocVQA Datasets and Their Evolution

The development of DocVQA datasets has closely mirrored the advancement in model capabilities. The seminal DocVQA dataset (Mathew et al., 2021) established foundational benchmarks for document understanding, focusing primarily on in-line questions where answers could be found within single text spans. This was followed by datasets that introduced additional complexity:

Single-Page Complex Reasoning: Datasets like CS-DVQA (Du et al., 2022) and RDVQA (Wu et al., 2022) pushed beyond simple text extraction by requiring commonsense reasoning and regional understanding. ArxivQA (Li et al., 2024b) further expanded the challenge by incorporating multiple-choice questions based on academic documents with mixed elements like tables, figures, and charts.

Multi-Page Understanding: The introduction of multi-page datasets marked a significant evolution in the field. SlideVQA (Tanaka et al., 2023) pioneered questions spanning multiple presentation slides, while MP-DocVQA (Tito et al., 2023) extended document coverage to up to 20 pages. DUDE (Van Landeghem et al., 2023) enriched the challenge by introducing diverse answer types, including lists and arithmetic problems. SPIQA (Pramanick et al.) specifically targeted academic content, requiring sophisticated understanding of scientific figures and plots.

Long-Form Document Understanding: As MLLMs demonstrated increasing capability in handling standard DocVQA tasks, more challenging benchmarks emerged. MMLongBench-Doc (Ma et al.) represents the current frontier, testing models' ability to reason over long-form documents with complex, multi-step questions. However, none of these datasets specifically target the challenge of locating and reasoning about minute details within larger document contexts—the gap our NiM-Benchmark aims to address.

A.1.3 Methods in Document Understanding

The methodological approach to document understanding has seen several paradigm shifts:

OCR and Layout-Aware Models: Early approaches relied heavily on OCR-based pipelines (Subramani et al., 2020), treating text and visual elements separately. The introduction of layout-aware models like LayoutLM and its variants (Xu et al., 2020b,a; Huang et al., 2022) marked a significant advance by incorporating spatial information and document structure into the modeling process.

End-to-End Multimodal Models: The emergence of powerful MLLMs (Team et al., 2023; Driess et al., 2023; Peng et al., 2023; OpenAI, 2023) has enabled end-to-end document understanding approaches. Recent methods like CREAM (Zhang et al., 2024b) and CFRET (Zhang et al., 2024a) have demonstrated strong performance across various DocVQA tasks.

Retrieval-Augmented Generation: For larger documents, retrieval-augmented generation (RAG) has emerged as a crucial technique. Methods like ColPali (Faysse et al., 2024b) and M3DocRAG (Cho et al., 2024a) have shown promise in efficiently handling large document collections. However, these approaches often process entire document regions without considering information granularity, leading to inefficiencies when answers lie in small, specific regions.

Benchmarks	# Pages/ Document	Unanswerable Questions	Granular Questions	Document Relevance	Answer Source	Domains
DocVQA (Mathew et al., 2021)	1	✗	✗	✗	TXT/L/C/TAB/I	Industry Docs
ChartQA (Masry et al., 2022)	1	✗	✗	✓	C	Statista, Pew, OWID, OECD
InfoVQA (Mathew et al., 2022)	1.2	✗	✗	✗	L/C/TAB/I	Infographics Browsing
TAT-DQA (Zhu et al., 2022)	1.1	✗	✗	✗	TXT/TAB	Finance Reports
DUDE (Van Landeghem et al., 2023)	5.7	✓	✗	✗	TXT/L/C/TAB/I	Books, Media, Public Docs
MP-DocVQA (Tito et al., 2023)	8.3	✗	✗	✗	TXT/L/C/TAB/I	Industry Docs
ArxiVQA (Li et al., 2024a)	1	✗	✗	✗	L/C/I	Scientific papers
SlideVQA (Tanaka et al., 2023)	20	✗	✗	✗	TXT/L/C/TAB/I	SlideDecks
MMLONGBENCH-DOC (Ma et al.)	47.5	✓	✗	✓	TXT/L/C/TAB/I	Research and Financial Reports, Academic Papers, Industry Files
NiM-Benchmark (Ours)	29	✓	✓	✓	TXT/L/C/TAB/I	Menus, Academic Papers, Magazines, Website SS, Lectures SS, Newspapers

Table 3: Comparison of benchmarks based on document-level attributes and question types. SS is Screenshots

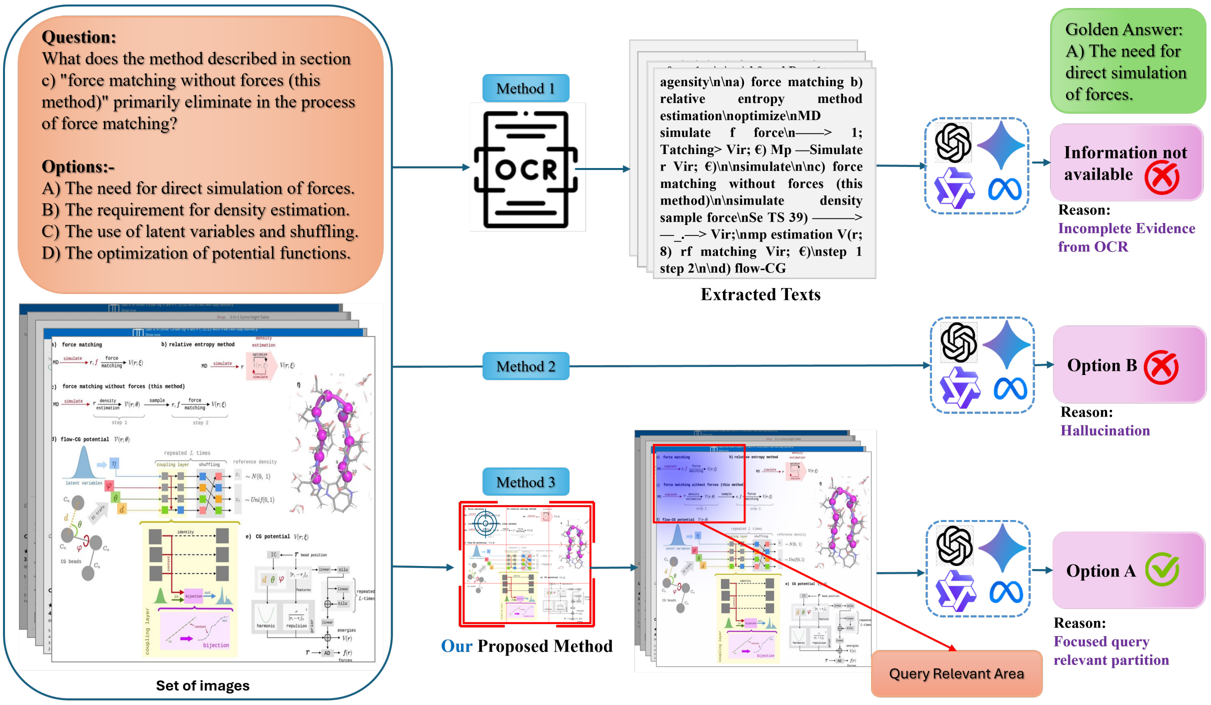


Figure 3: Spot-IT method comparison with existing methods. **We highlight failure cases of existing methods and illustrate how Spot-IT effectively overcomes these challenges.**

Figure 3 shows a comparison of our method, Spot-IT, with existing methods.

A.1.4 Fine-Grained Visual Analysis and Attention Mechanisms

While fine-grained visual analysis has been extensively studied in natural images, its application to documents presents unique challenges:

Visual Prompting: Recent work in visual prompting (Wu et al., 2024) has shown promising results in directing model attention. Techniques including bounding boxes (Lin et al., 2024), markers (Shtedritski et al., 2023), and pixel-level annotations (Yang et al., 2023) have proven effective in natural image understanding tasks.

Document-Specific Challenges: Documents present unique challenges for fine-grained analysis due to their hierarchical structure, complex layouts, and the need to preserve both spatial and semantic relationships. Our Spot-IT addresses these challenges through a novel question-guided attention mechanism that adapts visual prompting techniques specifically for document understanding tasks.

A.2 Additional Results

Table 5 presents our systematic optimization of the Spot-IT framework. In our comparison of CLIP and SigLIP for patch-query similarity, SigLIP consistently outperforms CLIP, achieving an accuracy of 0.59 compared to 0.56. Table 6 reports the effect

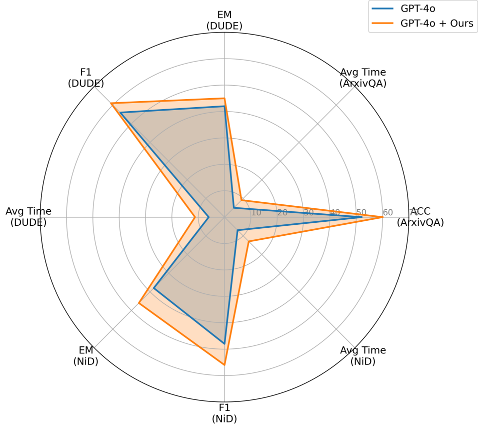


Figure 4: Accuracy and response time comparison of GPT-4o and GPT-4o + Ours on (a) ArxiVQA, (b) DUDE, and (c) NiM-Benchmark.

of varying the number of patches (N), showing that accuracy increases with N and peaks at N=7 (0.61), before slightly declining at N=8 (0.60). We select N=6 (0.59 accuracy) to ensure better generalizability, striking a balance between strong performance and avoiding potential overfitting at the peak.

Tables 4 present comprehensive evaluation results across multiple document understanding benchmarks. It shows consistent improvements with our method over GPT-4o across DocVQA, InfoVQA, and ChartQA, each evaluated on 200 representative questions. For long document understanding, MMlongbench-doc results—evaluated on 54 samples across two runs—further validate the effectiveness of our approach, showing improvements across all metrics: Exact Match, F1, and ANLS.[1]

A.3 Additional Figures and Tables

1. Table 7 provides a comprehensive overview of the NiM-Benchmark dataset. Table 8 lists the data sources used to construct NiM-Benchmark, while Table 9 outlines the distribution of question categories across various domains. This structured distribution ensures a balanced representation of domain-specific questions, enabling a thorough evaluation of model performance in diverse scenarios.

2. Table 10 presents results from a Turing test, comparing human-generated and machine-generated responses across different question categories. These results offer insights into the models' capability to generate responses that closely resemble human-like reasoning and linguistic patterns.

3. Figure 4 illustrates a comparative performance analysis between GPT-4o and its enhanced variant (GPT-4o + Ours) across multiple well-established benchmarks, including ArxiVQA, DUDE, and NiD-Benchmark. The results demonstrate that Spot-IT leads to a measurable improvement in accuracy across various tasks. However, this gain comes at the cost of slightly increased inference time, suggesting a trade-off between performance enhancement and computational efficiency.

4. Figure 5 provides an in-depth examination of the performance gap between AI models and human annotators on the NiD-Benchmark dataset across different domains. The analysis reveals that human responses consistently achieve superior F1 and EM (Exact Match) scores, while also exhibiting a longer average response time. This discrepancy underscores the limitations of existing AI models in achieving human-level comprehension and contextual reasoning, further motivating future advancements in model architectures and training paradigms.

[1]We used fewer samples for MMlongbench-doc due to the high computational cost associated with long documents. For similar reasons, we restricted our evaluation to the GPT-4o model across all datasets.

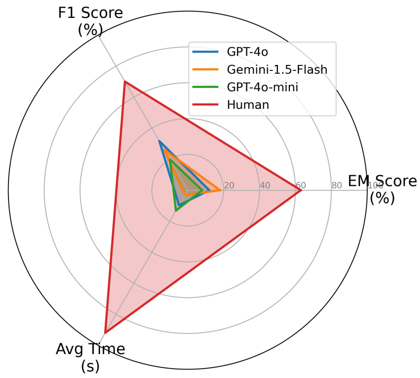


Figure 5: Accuracy and response time comparison on NiM-Benchmark (a) for GPT-4o, GPT-4o-mini, Gemini-1.5-Flash, and human.

A.4 Occlusion Sensitivity Analysis

MLLMs integrate both visual and textual modalities to answer queries about images. Understanding how these models focus on different parts of an image is crucial for interpretability. We implement an occlusion sensitivity method to identify critical image regions that affect model predictions.

Method	DocVQA			InfoVQA			ChartQA			MMlongbench-doc		
	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS	EM	F1	ANLS
GPT-4o	0.70	0.85	0.73	0.50	0.54	0.49	0.29	**0.32**	**0.29**	0.34	0.44	0.42
GPT-4o + Our Method	**0.73**	**0.88**	**0.74**	**0.52**	**0.56**	**0.50**	**0.30**	**0.32**	**0.29**	**0.37**	**0.48**	**0.45**

Table 4: Performance (EM / F1 / ANLS) on DocVQA, InfoVQA, ChartQA, and MMlongbench-doc datasets.

Model	ACC
GPT-4o	0.53
GPT-4o + Spot-IT(CLIP)	0.56
GPT-4o + Spot-IT(SigLIP)	**0.59**

Table 5: Accuracy scores for ArxivQA to compare Siglip and Clip for similarity matching of patch and query

Method	Acc
GPT-4o	0.53
Spot-IT + GPT-4o(N=3)	0.60
Spot-IT + GPT-4o(N=4)	0.58
Spot-IT + GPT-4o(N=5)	0.59
Spot-IT + GPT-4o(N=6)	0.59
Spot-IT + GPT-4o(N=7)	**0.61**
Spot-IT + GPT-4o(N=8)	0.60

Table 6: Effect of Number of Patches (N) on Accuracy Score for ArxivQA

A.4.1 Model and Dataset

The Qwen2-VL model (Wang et al., 2024) is employed for answering image-based queries. The dataset used is the ArxiVQA dataset..

A.4.2 Occlusion Sensitivity Analysis

Given an image I of size (W, H) and a query Q, we systematically occlude square patches of the image and measure the change in response probability. The procedure is as follows:

1. Compute the model's original response probability P_{orig}.
2. Slide an occlusion window of size $S \times S$ with stride T over the image.
3. Replace the windowed region with a neutral color (e.g., black or gray).
4. Compute the new response probability P_{occ} after occlusion.
5. Compute the sensitivity score as:
$$S(x, y) = P_{orig} - P_{occ} \tag{12}$$
where (x, y) are the coordinates of the occluded patch.
6. Generate a heatmap from $S(x, y)$ values and apply Gaussian smoothing.

A.4.3 Probability Calculation

To determine the probability of a model's response, the output logits are converted into probabilities using the softmax function:

$$P(y) = \frac{e^{z_y}}{\sum_i e^{z_i}} \tag{13}$$

where z_y is the logit corresponding to the generated response.

A.5 Extended Spot-IT Error Analysis

We analyze our method on ArxivQA using GPT-4o on 500 samples, where 200 samples were incorrect. We randomly selected 50% of these samples and categorized the errors as follows:

- **Dataset Error (19 cases)**: The dataset had 14 cases of incorrect or ambiguous ground-truth answers, and some questions lacked the necessary context, leading to unavoidable evaluation errors.
- **Retrieval Error (22 cases)**: The retrieval module (Faysse et al., 2024a) failed to fetch relevant information, leading to incorrect answers.
- **Patch Formation (25 cases)**: The patch was incorrectly formed due to a static grid size, leading to improper image cropping and loss of answer context, which caused incorrect matching with the query.
- **Patch Selection (26 cases)**: Incorrect semantic similarity matching occurred between the patch and the input query due to the query's complexity.
- **LLM Fault (8 cases)** Despite having the correct patched image, the Large Language Model sometimes fails to provide the correct answer, particularly for complex questions.

A.6 Extended NiM-Benchmark Error Analysis

We evaluate the performance of NiM-Benchmark on GPT-4o by randomly selecting 20 samples from

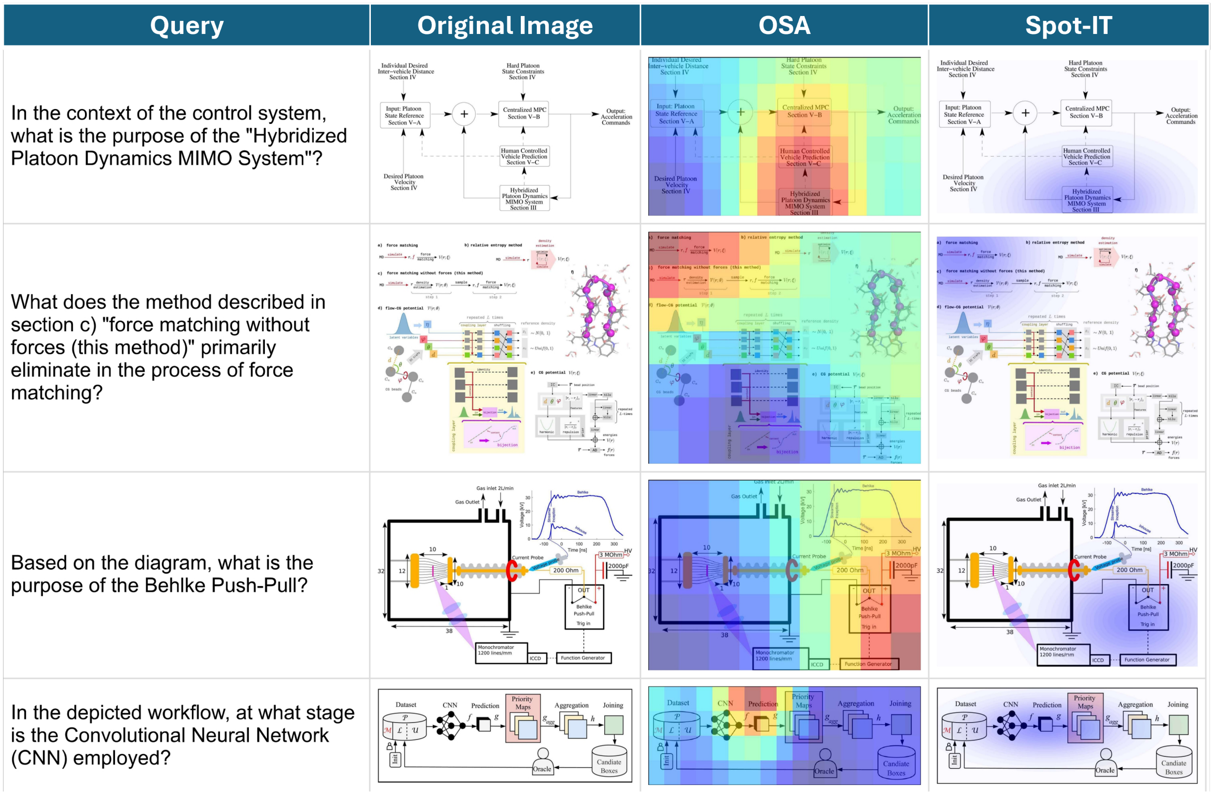


Figure 6: Occlusion Sensitivity Analysis(OSA) comparison with Spot-IT. **Demonstrating the correlation between where the MMLLM searches for the answer and where Spot-IT highlights the images to assist MMLLMs.**

all 6 domains domain and categorized them as follows:

- **Incomplete Evidence (47 cases)**: MLLM is not able to find an evidence to answer the question.
- **Hallucinated Evidence (28 cases)**:MLLM is either answering unanswerable questions or hallucinating the response.
- **Perceptual Error (24 cases)**: MLLMs struggle to perceive details such as incorrect decimal placements, leading to inaccurate answers.
- **Reasoning Error (25 cases)**: MLLMs struggle to reason accurately, often selecting the first piece of evidence in the relevant section without verifying its correctness.
- **Irrelevant Answer (5 cases)**: MLLM is not able to reason deeply and relies on pattern matching, leading to irrelevant answers. It often prioritizes the most prominent or recent context, resulting in inaccurate responses.
- **Knowledge Lacking (1 case)**: MLLMs may lack knowledge due to outdated training data, insufficient domain-specific information, or limited context understanding. Additionally, they may struggle with complex reasoning or nuanced details not well-represented in the training corpus.

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

Table 7: Dataset Statistics for NiM-Benchmark

A.7 Sample Illustrations from NiM-Benchmark

Table 11 represents examples from NiM-Benchmark encompassing multiple domains and categories to support diverse research applications. The dataset integrates visually rich images from domains such as website screenshots, lecture slides, restaurant menus, magazines, newspapers, and research papers. Each instance is categorized into Boolean, unanswerable, common sense, reasoning, comparative, and inline question-answering tasks.

Domain	Source
Restaurant Menus	Various Sources including Heathrow Restaurants, London Stansted Restaurants etc.
Academic Papers	Arxiv (2024-2025)
Magazines	freemagazines.top
Newspapers	Times of India, The Hindu, Hindustan Times (2024-2025)
Website Screenshots	CoVA dataset (Kumar et al., 2022)
Lecture Screenshots	MIT 6.034 AI, Fall 2010 (MIT OCW)

Table 8: Data Sources used to construct the NiM-Benchmark dataset across different domains

Domain	Count	Domain	Count	Domain	Count
News Paper		**Lectures**		**Screenshots**	
Inline	199	Inline	48	Inline	203
Comparative	10	Comparative	–	Comparative	–
Unanswerable	7	Unanswerable	15	Unanswerable	3
Reasoning	–	Reasoning	25	Reasoning	35
Boolean	–	Boolean	12	Boolean	5
Commonsense	–	Commonsense	2	Commonsense	–
Total	**216**	**Total**	**102**	**Total**	**246**
Academic Paper		**Magazines**		**Menus**	
Inline	185	Inline	180	Inline	143
Comparative	22	Comparative	9	Comparative	21
Unanswerable	8	Unanswerable	3	Unanswerable	–
Reasoning	5	Reasoning	10	Reasoning	–
Boolean	–	Boolean	–	Boolean	23
Commonsense	–	Commonsense	–	Commonsense	7
Total	**220**	**Total**	**202**	**Total**	**194**

Table 9: NiM-Benchmark Distribution of Question Categories Across Domains

A.8 Spot-IT Qualitative Examples

Figures 7, 8, and 9 present qualitative examples from the NiM benchmark, demonstrating its applicability across diverse domains such as restaurant menus, website screenshots, and lecture slides. These examples emphasize how NiM focuses on fine-grained visual question answering, requiring models to reason over localized and domain-specific visual details. Furthermore, the effectiveness of the proposed Spot-IT method is highlighted, as it successfully identifies and highlights the query-relevant regions in each image. By drawing attention to the most informative parts of the visual input, Spot-IT facilitates better grounding for multimodal large language models, thereby improving their interpretability and VQA performance across different real-world document types.

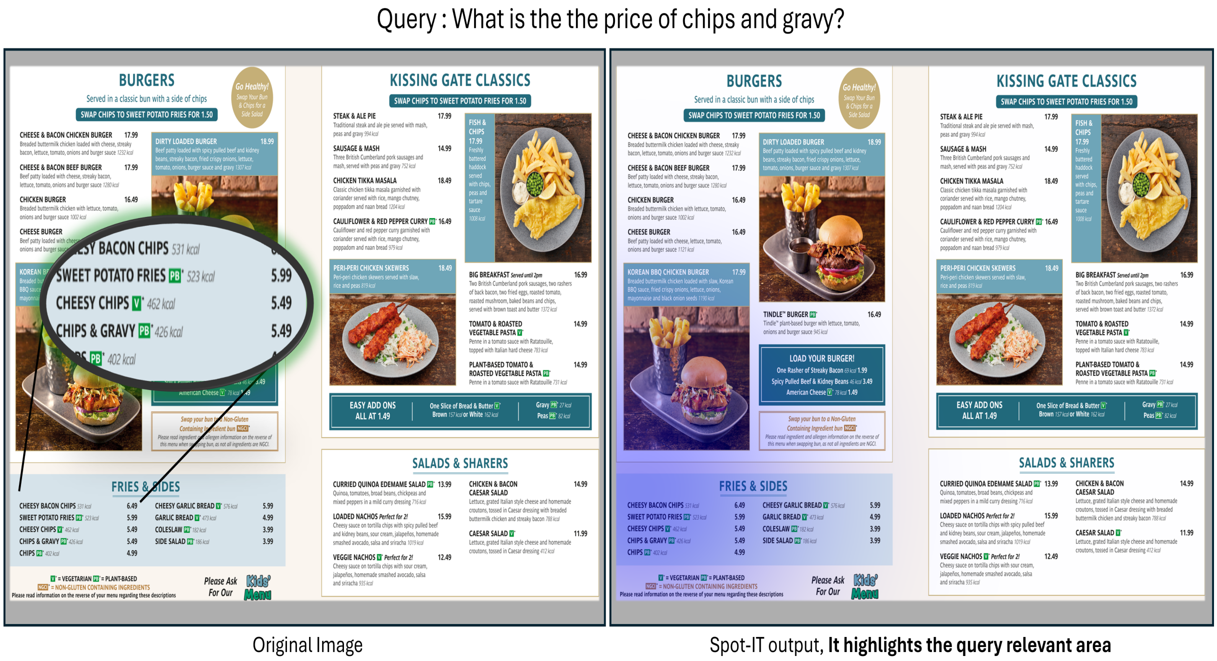


Figure 7: An example from the restaurant menu domain in the NiM benchmark. The Spot-IT method accurately highlights the query-relevant region.

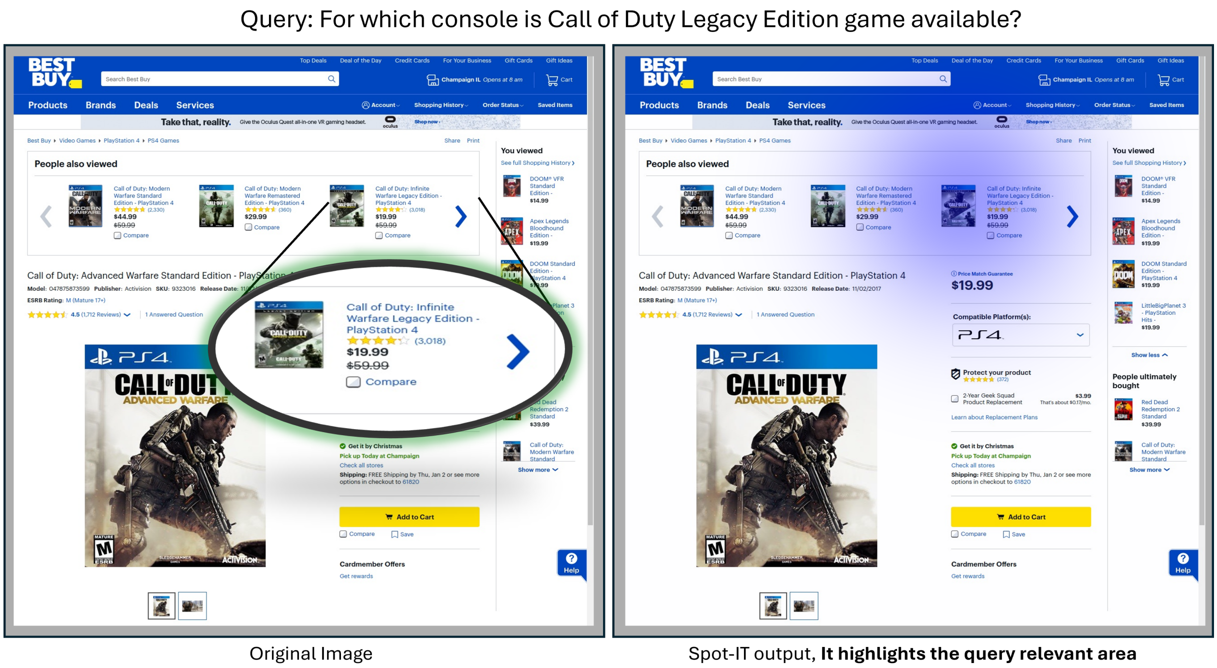


Figure 8: An example from a website screenshot in the NiM benchmark.Spot-IT successfully localizes the visual region relevant to the query.

Ground Truth	Gemini 2.0 Flash		Human verifier 1	
	Predicted Human	Predicted Machine	Predicted Human	Predicted Machine
Human	146	54	170	30
Machine	143	57	160	40
Total	289	111	330	70
Ground Truth	**Claude 3.5 Sonnet**		**Human verifier 2**	
	Predicted Human	**Predicted Machine**	**Predicted Human**	**Predicted Machine**
Human	181	19	171	29
Machine	176	24	162	38
Total	357	43	333	76

Table 10: **Turing Test and LLM as a Judge Results.** We find that the generated questions in our NiM-Benchmark are classified as human-generated with a moderately high agreement score

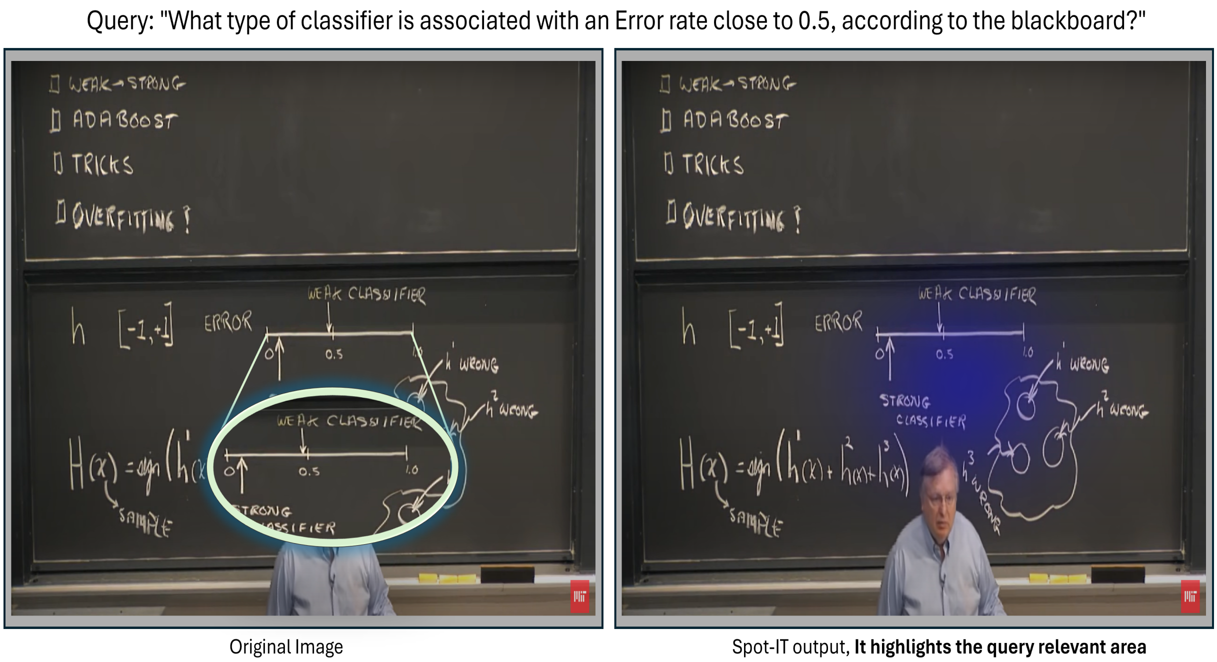


Figure 9: An example from the lecture screen shot domain in the NiM benchmark. Spot-IT effectively highlights the query-specific portion of the blackboard.

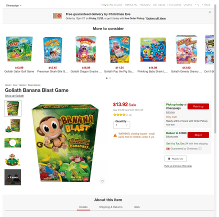
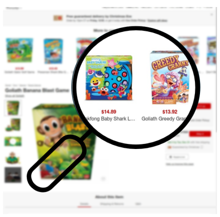
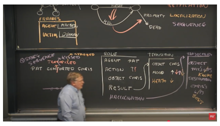
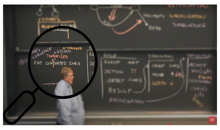
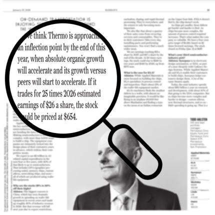
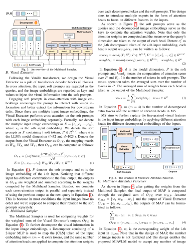
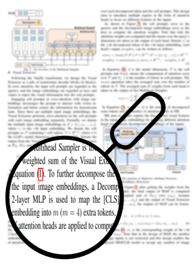
Domain	Category	Image	Region of Interest	Question	Answer
Website Screen Shot	Boolean			The game "Greedy Granny" and "Baby Shark" are priced the same (True/-False)?	False
Lecture Screen Shot	Unanswerable			Who Hugged Chris?	Information not available
Restaurant Menus	Common Sense			Is the Nawarattan Korma dish vegetarian?	Yes
Magazines	Reasoning			What is the estimated price of Thermo's stock if it trades at 25 times 2026 earnings?	$654
News Papers	Comparative			What was the record low value of the rupee against the dollar?	85.07
Research Papers	Inline			What is the value of m in the Decomposer's MLP?	4

Table 11: **Sample Illustrations from NiM-Benchmark.** Question-answer pairs across different domains, including the question, required context, question category, and relevant region of interest to find the answer.

A.9 All LLM Prompts Used for Evaluation and Dataset Generation

A.9.1 Prompt for Document VQA Evaluation

This prompt assesses a model's ability to answer questions based solely on document images, without external knowledge. Responses should be concise (preferably a single word or number). If the information is unavailable, the model should respond with `"Information not available."`

A.9.2 Customized Prompt for Document VQA Evaluation

This variant prioritizes information in blue-highlighted regions, considering the entire image only if necessary. Constraints on external knowledge, concise responses, and handling of missing information remain unchanged.

A.9.3 QA Generation Prompt for NiM-Benchmark

This prompt generates precise, challenging questions from document images. Each question should be natural, answerable from a small document portion, and uniquely identifiable. Necessary context must be explicit, avoiding vague references.

Only 2–3 high-quality questions per document should be produced; otherwise, output `"NA."` The output follows a structured JSON format for consistent benchmarking.

Prompt for Document VQA Evaluation (Ma et al.)

Task:

[Images]

`Read the above Images and answer this question`

Instructions:

- `DO NOT use external knowledge.`
- `Provide a one-word or numerical answer if possible.`
- `If information is unavailable, state "Information not available."`

Customized Prompt for Document VQA(for Spot-IT) Evaluation

Task:

[Images]

`Read the above Images and answer this question`

`Focus on the BLUE Highlighted area in images as it is more relevant to the query. First, try to answer only using the highlighted area, and if not found, then, consider whole image`

Instructions:

- `DO NOT use external knowledge.`
- `Provide a one-word or numerical answer if possible.`
- `If information is unavailable, state "Information not available."`

QA Generation Prompt for NiM-Benchmark Benchmark

Task:

[Images]

`You are very good in question making from documents. I am giving you a task to make some questions from some pages from a document.`

Instructions:

- The questions should be precise. Each question should be answerable from a very small portion of the document and relevant to the textual and visual elements of the provided image.
- Questions should be natural and easy to understand. yet,questions should be challenging enough that even you would find them difficult to answer immediately.
- Ensure the questions are open-domain so that even if multiple documents are provided, the question remains uniquely identifiable and answerable.
- Include all necessary information to make the question unique and answerable. Avoid vague references like "according to the given article" or "mentioned in the article". Explicitly include the full information if needed.
- Create only 2-3 high-quality questions. If a quality question cannot be made, return "NA". However, ensure that effort is made to create a good question.
- **Accepted Questions:**
  - "Question": "Who accused AAP of supporting 'terrorist sympathizers' during Punjab elections?"
  "Answer":Anurag Thakur"
  - "Question": "What was the altitude of Sandakphu where the tourist died?"
  "Answer": "11,900 feet"
- **Rejected Questions:**
  - "Question": "Who is the alleged associate of Partha Chatterjee mentioned in the article?"
  Don't make such questions that reference the artcile.
  - "Question": "Which company is prominent in biodiversity monitoring using AI?"
  Such question is not acceptable because it is document specific. There can be multiple answers.
- Stick to the above format. If you are unable to create quality questions, return NA.

Output Format (JSON):

```
{
    "questions": [
        {
            "question": "the question",
            "answer": "the answer"
        },
        ...
    ]
}
```