VISA: Retrieval Augmented Generation with Visual Source Attribution

Anonymous ACL submission

Abstract

Generation with source attribution is important for enhancing the verifiability of retrieval-augmented generation (RAG) systems. However, existing approaches in RAG primarily link generated content to document-level references, making it challenging for users to locate evidence among multiple content-rich retrieved documents. To address this challenge, we propose *Retrieval-Augmented Generation with Visual Source Attribution* (VISA), a novel approach that combines answer generation with visual source attribution. Leveraging large vision-language models (VLMs), VISA identifies the evidence and highlights the exact regions that support the generated answers with bounding boxes in the retrieved document screenshots. To evaluate its effectiveness, we curated two datasets: Wiki-VISA, based on crawled Wikipedia webpage screenshots, and Paper-VISA, derived from PubLayNet and tailored to the medical domain. Experimental results demonstrate the effectiveness of VISA for visual source attribution on documents' original look, as well as highlighting the challenges for improvement. Code, data, and model checkpoints will be released.

1 Introduction

Retrieval-augmented generation (RAG) has become a key technique for enhancing the reliability in information-seeking processes (Gao et al., 2024). Traditional RAG pipeline directly generates an answer to a user query from retrieved candidate documents (Chen et al., 2017; Lewis et al., 2020). Yet, it is hard for users to verify the sources and appropriately trust generated answers, given that models could produce hallucinated content (Min et al., 2023; Malaviya et al., 2024). Recent works have introduced the generation with citation paradigm (Gao et al., 2023; Ye et al., 2024), prompting the model to not only generate answers but also directly cite the identifiers of the source documents. Such source attribution approaches make it possible for users to check the reliability of the outputs (Asai et al., 2024).

However, text-based generation with source attribution faces several issues: First, citing the source at the document level could impose a heavy cognitive burden on users (Foster, 1979; Sweller, 2011), where users often struggle to locate the core evidence at the section or passage level within the dense and multi-page document. Despite such granularity mismatch could be addressed through passage-citation-based generation methods — linking answers to specific text chunks, it requires non-trivial extra engineering efforts to match the chunk in the document source. Moreover, visually highlighting text chunks in the source document is more intuitive for users, but it remains challenging as it requires control over document rendering, which is not always accessible, such as in PDF scenarios.

Inspired by the recent document screenshot embedding retrieval paradigm — dropping the document processing module and directly using VLM to preserve the content integrity and encoding document screenshots for retrieval (Ma et al., 2024), we ask whether source attribution can also be integrated into such a unified visual paradigm to establish a fully visual, end-to-end verifiable RAG pipeline that is both user-friendly and effective?

To this end, we propose *Retrieval Augmented Generation with Visual Source Attribution* (VISA). In our approach, a large vision-language model (VLM) processes single or multiple retrieved document images and not only generates an answer to the user query but also returns the bounding box of the relevant region within the evidence document. As Figure 1 illustrated, this method enables direct attribution by visually pinpointing the exact position within the document, allowing users to quickly check the supporting evidence within the original context for the generated answer. VLMs are not restricted by document format or render-

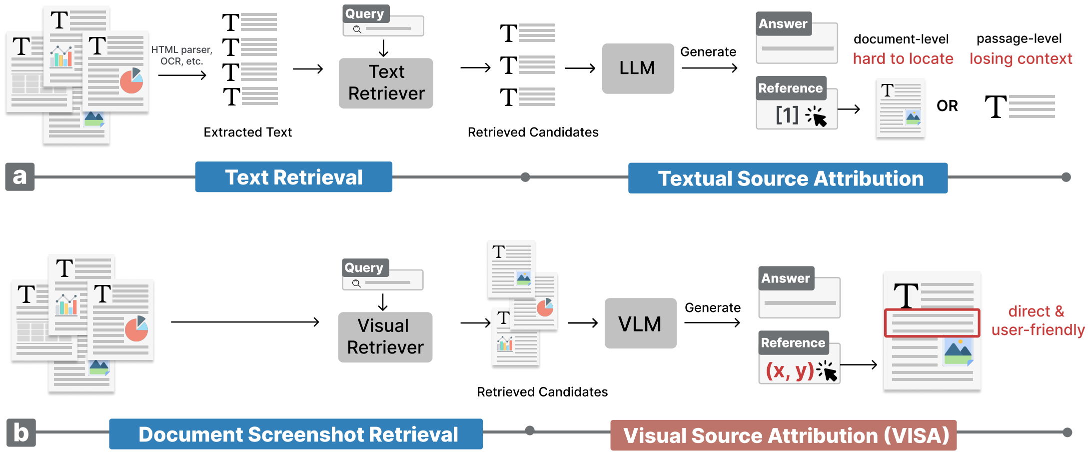


Figure 1: Comparison between (a) Text-based generation with source attribution in a RAG pipeline. and (b) Visual-based generation with source attribution in a V-RAG pipeline. VISA directly pinpoints the source evidence of the answer for user query in the original document with a bounding box.

ing, making them more versatile for diverse use cases. Moreover, this task serves as a meaningful evaluation of VLMs, assessing their ability to provide self-explanations and accurately localize supporting information within their input in an RAG paradigm. To the best of our knowledge, this is the first work to enable the visual source attribution in an end-to-end RAG framework using VLM.

To train and evaluate VISA, we curated two datasets: Wiki-VISA and Paper-VISA. Wiki-VISA is derived from the Natural Questions dataset (Kwiatkowski et al., 2019). It reconstructs the original Wikipedia webpages, using short answers as generation targets and corresponding long answer's HTML bounding box as source attribution targets. This dataset supports the test of model's ability to attribute sources across multi-document, multi-page, and multi-modal content. On the other hand, Paper-VISA, built from PubLayNet (Zhong et al., 2019) with synthetic query generation, focuses on the biomedical domain by evaluating performance on multi-modal scientific paper PDFs. Together, they provide diverse and challenging benchmarks for assessing the granularity and accuracy of source attribution in RAG systems. Our experiments, spanning both in-domain training and zero-shot evaluation, revealed existing state-of-the-art models like QWen2-VL-72B (Wang et al., 2024) struggle with precise visual source attribution in zero-shot prompting. Fine-tuning VISA on our curated datasets significantly improved model performance in visual attribution accuracy. Further analysis highlights key areas for improvement, such as enhancing bounding box precision for long image documents, multi-documents, and zero-shot generalization capabilities.

2 Related Work

2.1 RAG attribution

Open-domain question answering with LLMs often suffer from two key issues: hallucinations and outdated internal knowledge. Retrieval-Augmented Generation (RAG) has been recognized as an effective solution to these problems (Lewis et al., 2020; Gao et al., 2024; Ovadia et al., 2024). In RAG, relevant documents are first retrieved from an external database and then fed into LLMs alongside the question. This allows LLMs to reference the retrieved documents during answer generation. Furthermore, RAG can generate a list of citations attached to the generated answers, linking them to the retrieved documents so users can verify the accuracy of the output. This process is known as source attribution (Rashkin et al., 2023; Bohnet et al., 2023; Khalifa et al., 2024).

Typically, RAG with source attribution follows a text-only pipeline where all inputs and outputs, such as questions, retrieved documents, generated answers, and citations, are in textual form. Recently, vision-based RAG pipelines have emerged, where the retrieved documents are represented as screenshot images (Ma et al., 2024; Faysse et al., 2024), and VLMs process both textual questions and these document images to generate answers (Riedler and Langer, 2024; Xia et al., 2024; Yu et al., 2024; Cho et al., 2024). Compared to traditional text-only RAG, vision-based RAG can leverage structured and visual information from documents, such as tables, graphs, and images, which are often challenging to extract through text-only pipelines.

Our VISA attribution method proposed in this paper is a novel approach for vision-based RAG pipelines: directly drawing bounding boxes around the content in retrieved document screenshots that potentially supports the generated answers. This approach differs from existing attribution methods in two ways: (1) Granularity: Existing attribution methods often operate at the document level, requiring users to read entire documents to locate supportive content. In contrast, our method directly attributes the answer to specific content within the document, such as a passage, table, or image in the screenshot. (2) Presentation: Traditional attribution methods provide a list of textual citations, whereas our method uses bounding boxes, offering a visually-oriented form of attribution. This can help users quickly locate the relevant information.

2.2 Bounding Box Drawing with VLM

Bounding box-based object detection is a well-established task in computer vision (CV) (Zhao et al., 2019; Zou et al., 2023). Traditional approaches rely on convolutional neural networks (CNNs) (LeCun et al., 2015) or Vision Transformers (ViTs) (Dosovitskiy et al., 2021) to extract features and predict bounding boxes alongside object classification (Ren et al., 2015; Dai et al., 2016; Redmon et al., 2016; Carion et al., 2020).

Recent vision-language models (VLMs) like GPT4o (OpenAI, 2024), QWen2-VL (Wang et al., 2024), and PaliGemma (Steiner et al., 2024) have shown the ability to generate bounding box coordinates in an image-to-text manner, taking input images and generate the top-left and bottom-right coordinates of target objects. Methods like BuboGPT (Zhao et al., 2023) and GLAMM (Rasheed et al., 2024) integrate additional modules or modify the VLM architecture tailoring for the visual grounding tasks. Unlike traditional object detection or grounding that focuses on natural images, our method applies bounding box drawing to text-intensive document screenshots. In addition, we intentionally leave the VLM architecture unchanged, envisioning visual attribution eventually can be naturally integrated into general-purpose VLM training data.

Grounding elements on screenshots have been explored in GUI agent systems (Cheng et al., 2024; Lin et al., 2024), where bounding boxes are used to localize UI elements like buttons. However, these approaches focus on GUI contexts, our work targets visual source attribution in vision-based RAG processes, grounding bounding boxes to locate evidence within document images.

3 Method

3.1 Task Definition

Our model VISA is a novel source attribution method primarily designed for vision-based RAG systems. To formally define the task of RAG with visual-based source attribution: given a textual user query q as the RAG system input, the retrieval component of the system needs to retrieve a set of candidate documents $D = \{d_1, ..., d_n\}$ from corpus $\mathcal{C}$. Then the generation component of the system needs to return three outputs: an answer a that answers the query q, the identifier i of the most relevant document d^* in D, and a bounding box coordinates $B_{d^*} = [(x_1, y_1), (x_2, y_2)]$ within d_* that highlight the content supporting the generated answer a.

In a vision-based RAG setup, user queries are textual, while all documents in the corpus $\mathcal{C}$ are screenshots of documents (e.g., webpages or PDF pages) provided as image inputs.

3.2 Generation with Visual Source Attribution

This paper focuses on VISA within the generation component of vision-based RAG systems. As discussed in the previous section, VISA must handle multimodal input. To achieve this, we leverage VLMs for implementing VISA. Specifically, for a given query and a set of retrieved candidate documents (i.e., screenshots of documents), the system processes the inputs as follows: query tokens are directly input into the language model, while document screenshots are first processed by the image encoder to extract image representations, which are then fed into the language model.

The language model subsequently generates the answer, the identifier of the relevant document, and the xy-coordinates of the bounding box's top-left and bottom-right corner on the content that supports the generated answer. Notably, this entire process can be framed as a next-token prediction task. Finally, the generated identifier and bounding box coordinates are used to draw the bounding box on the target document screenshot, which is presented to the user along with the generated answer.

Technically, existing instruction-tuned VLMs, such as Qwen2-VL-72B (Wang et al., 2024), can potentially be prompted to perform VISA in a zero-shot manner. However, we find that VISA remains a challenging task. Consequently, further supervised fine-tuning on a dedicated VISA task dataset is necessary. In the next section, we introduce the datasets we crafted specifically for training and evaluating VISA.

3.3 Dataset Acquisition

The training and evaluation data suitable for the VISA task needs to be formatted as follows: the input consists of a textual query and document screenshot images as multimodal inputs, while the target outputs include the textual short answer, the relevant document identifier, and the coordinates of the bounding box. To create datasets that meet these requirements, we craft existing publicly available datasets to support the training and evaluation of our proposed VISA method.

Wiki-VISA is derived from the Natural Questions (NQ) dataset (Kwiatkowski et al., 2019). The original NQ dataset provides natural questions, along with short and long answers sourced from Wikipedia webpages. We use the short answers as answer targets. However, the original dataset does not contain the original webpage screenshots. We use the Selenium Python toolkit[1] to access and render the webpage with the original URL with a history version stamp. And take a screenshot with 980 pixels width and up to 3920 pixels (4 pages) height. Using the long answer, we identify the corresponding element in the HTML from which the long answer is derived. We then draw a bounding box around this element to obtain the coordinates. Notably, the answers in this dataset can come from various elements, such as passages, tables, lists, or images within the webpage. Since the questions and answers in Wiki-VISA are human-judged, we consider this dataset a high-quality, supervised dataset and evaluation for VISA on general knowledge, with Wikipedia webpage.

Paper-VISA is derived from PubLayNet (Zhong et al., 2019), a dataset originally designed for document layout analysis of single-page PubMed PDF documents (adhering to a 3:2 aspect ratio). PubLayNet provides bounding box coordinates and class labels (e.g., title, text, table, figure, etc.) for each element in a paper's PDF screenshot. However, the dataset does not include queries or answers associated with each document. To address this limitation, we leverage instruction-tuned VLMs (e.g. Qwen2-VL-72B) to synthetically generate queries and answers. Specifically, for each paper screenshot sample in the PubLayNet training data, we select a bounding box within the sample and overlay it on the screenshot. The modified screenshot is then input to the VLM with a prompt designed to instruct the model to generate a question and a short answer based on the content within the bounding box. See Appendix A.10 for the prompt details and generation example. By augmenting the original PubLayNet in this way, we create synthetic queries and answers, enabling it to support VISA training. We consider the resulting Paper-VISA dataset as synthetic training and evaluation for scientific paper PDFs in the medical domain.

FineWeb-VISA is based on the FineWeb-edu corpus (Penedo et al., 2024), a high-quality text corpus of crawled webpages. We sampled 60k webpage URLs and used Selenium to capture screenshots of diverse, content-rich webpages (in 980x3920 pixels). A passage containing more than 50 words was randomly selected as the target source. A bounding box was drawn around the selected content, and a VLM was prompted to generate a query and short answer supported by the target content, similar as Paper-VISA. Although Fineweb-VISA provides a diverse layout, it does not guarantee be high-quality data as human annotated in Wiki-VISA or Paper-VISA that assessing a specific domain, we only leverage Fineweb-VISA as training data to analysis zero-shot and data augmentation effectiveness.

3.4 Multi-Candidates

By now, each query is paired with the triplet of a positive document, target short answer, and target evidence bounding box. To set up a RAG experimental environment for evaluating VISA, we in addition need to let the generator take multiple candidates as input, simulating the scenario that the generator is taking multiple retrieval candidates and attributing the evidence in most relevant documents. Given the query q, we use a retriever R to retrieve top-k candidates. And randomly sampled $m-1$ candidates that are not ground truth as hard negative candidates. The hard negative candidates are mixed with the one ground truth document together as the input for the multi-document VISA. The reason we did not directly take top-m documents as the retrieval candidate is that we do not

[1] https://pypi.org/project/selenium/

Dataset	# Train	# Test
Wiki-VISA	87k	3,000
Paper-VISA	100k	2,160
Fineweb-VISA	60k	-

Table 1: Datasets statistics for train and test splits.

want VISA biased on a specific retriever and position of the candidate docs. Generally, our model VISA does not rely on the type of retriever. It can be either a traditional text-based retriever that indexes the document with extracted text or a recent document screenshot retriever that directly indexes the original document screenshot. However, integrating with those visual-based retrievers enables us to build an end-to-end RAG solution without the necessity of explicit document content processes such as HTML parsing or OCR. Thus, we leverage an off-the-shelf Document Screenshot Embedding (DSE) model (Ma et al., 2024) to serve as the retrieval component of the RAG system. When encoding queries and documents, the model directly encodes textual queries and document screenshot images into single vector embeddings and performs cosine similarity search during inference. In this work, we set $k = 20$ and $m = 3$.

Additionally, an RAG pipeline may have the chance of having no ground truth document returned from the retriever. We use a probability of 20% to randomly replace the ground truth document in the candidates, to access the model's capability to detect no-answer situations. After these operations, the data statistics are shown in Table 1.

4 Experiment Setup

4.1 Evaluation

Evaluation metrics assessed both generated answers and bounding box predictions. Relaxed exact match (EM) was used to measure generated answer accuracy, considering a generated answer correct if it shares a subsequence relationship with the golden answer and differs by no more than 20 characters. Intersection over Union (IoU) was calculated to determine bounding box precision, with an IoU threshold of 0.5 indicating a correct prediction.

To analyze performance across varying content types, test samples were categorized by the modality and location of the evidence. For Wiki-VISA, categories included first-page passages, passages beyond the first page, and non-passage content such as tables and figures. For Paper-VISA, since it is a single-page document, categories were divided into passage and non-passage content. The overall accuracy for each dataset was computed as a macro average across these categories.

We evaluate the effectiveness of VISA in two different settings: *Single oracle candidate* and *Multi-candidate*. *Single oracle candidate* setting solely evaluates the generation and visual attribution component. We conduct controlled experiments by training and testing the VLMs using only a single ground truth relevant document screenshot as input. In this setup, it is guaranteed that the answer can be found within the input document. The VLMs do not need to predict the relevant document identifier and can focus exclusively on answer generation and bounding box prediction.

In a *Multi-candidate* setting, the model is evaluated on its ability to distinguish relevant documents from irrelevant ones, in addition to generating accurate answers and bounding boxes. This setup better reflects the RAG scenarios in which multiple candidate documents are retrieved, and the model must not only generate a correct response but also attribute it to the correct source document. For the *Multi-candidate* evaluation, we assess two configurations: *Multi-candidate, Oracle in Candidates* which has ground truth in candidates, this setting has the same query set as the single setting, hence directly comparable. *Multi-candidate, Oracle Not in Candidates* evaluated on the queries with no ground truth documents in candidates, assessing the model's ability to recognize when there is no supporting evidence

4.2 Training Details

To train vision-language models (VLMs) for answer generation with VISA, we initialized the models using the open-source Qwen2-VL-2B-Instruct and Qwen2-VL-7B-Instruct (Wang et al., 2024), finetuning on training datasets (Sec. 3.3).

We first trained the models in a single-candidate setup, where the input was limited to a single oracle document image. In this setup, the model was trained to generate both the answer and its corresponding bounding box. We used the prompt template provided in Appendix A.8 to format the model's input and output. Next, we trained the models in a multi-candidate setup. Here, the model received three document candidates and the task was to generate the identifier of the relevant document (if present), the answer, and the bounding box for the evidence. For cases where no relevant document was present (20% of the training samples),

Method	Wiki-VISA								Paper-VISA					
	Average		[<1] Passage		[>1] Passage		Non-Passage		Average		Passage		Non-Passage	
	bbx	ans	bbx	ans	bbx	ans	bbx	ans	bbx	ans	bbx	ans	bbx	ans
Zeroshot Prompt, Single Oracle Candidates														
QWen2-VL-72B	1.5	60.4	3.4	58.5	0.1	54.9	0.9	67.9	1.5	43.1	0.5	40.2	2.5	45.9
Fine-tune, Single Oracle Candidates														
VISA-2B-single	37.5	57.1	70.0	61.1	18.7	44.9	23.8	65.3	63.0	38.3	50.6	34.4	75.3	42.1
VISA-7B-single	54.2	65.2	75.6	66.5	50.1	56.0	36.8	73.1	68.2	43.8	58.1	41.6	78.2	45.9
Fine-tune, Multi Candidates, Oracle in Candidates														
VISA-2B-multi	22.5	37.9	46.5	46.1	6.4	27.2	14.6	40.5	51.3	32.4	41.1	30.1	61.4	34.7
VISA-7B-multi	32.3	41.8	51.7	48.6	23.0	32.7	22.2	44.1	59.9	39.2	47.7	35.9	72.0	42.4
Fine-tune, Multi Candidates, Oracle Not in Candidates														
VISA-2B-multi	73.7	84.9	68.0	82.0	73.2	84.9	80.0	87.7	95.2	95.2	97.2	97.2	93.1	93.1
VISA-7B-multi	82.2	91.0	75.1	87.6	84.0	91.4	87.4	94.0	95.6	95.6	97.2	97.2	93.9	93.9

Table 2: Effectiveness of VISA on Wiki-VISA and Paper-VISA datasets for bounding box accuracy (bbx) and answer accuracy (ans). Fine-tuned models are trained individually on in-domain data. The *Multi-Candidate, Oracle in Candidates* setting uses the same query set as the Single Oracle Candidates setting, allowing direct comparison. The Oracle Not in Candidates setting is evaluated on the queries with no ground truth documents in candidates.

the model was trained to generate "No answer." We used the prompt template in Appendix A.9 to format the model's input and output.

During the training, random cropping was applied outside of the bounding box. This augmentation exposed the model to varying input sizes, which enhanced its zero-shot effectiveness on unseen document layouts. Bounding box targets were represented using absolute coordinate values. We also explored normalizing the scale of bounding box coordinates to values in the range[0-1]. Details can be found in Appendix A.3 and A.4.

5 Experimental Results

Table 2 presents the performance of VISA on the Wiki-VISA and Paper-VISA datasets across different experimental settings. Zero-shot prompting results reveal the difficulty of directly applying state-of-the-art VLMs to the visual source attribution task. QWen2-VL-72B achieves a reasonable answer generation accuracy of 60.4% on average on Wiki-VISA but fails to deliver effective bounding box predictions, with only 1.5% accuracy. This observation is consistent on Paper-VISA. These highlight the limitations of existing VLMs in pinpointing the source evidence in original documents with proper location and granularity.

Fine-tuning on our crafted training data enables the model to effectively execute the task. In the single-candidate setup, where the model processes only the relevant document, fine-tuned models demonstrate substantial gains compared to zero-shot prompting a much larger model. On Wiki-VISA, the 7B variant achieves 54.2% bounding box accuracy and 65.2% answer accuracy, while on Paper-VISA, the corresponding scores reach 68.2% and 43.8%. It further demonstrates that the effectiveness of VISA is influenced by document characteristics, such as content location and modality. For Wiki-VISA, bounding box accuracy is significantly higher for passages on the first page ([<1] passage) compared to passages beyond the first page ([>1] passage). For example, the 2B variant achieves 70.0% accuracy for [<1] passages but only 18.7% for [>1] passages, indicating the challenges posed by long, multi-page documents. The larger model, the 7B variant, narrows this gap, reflecting the better handling of long-context inputs. Non-passage content, such as tables and figures, also have obviously a different level of grounding effectiveness, indicating the difference of effectiveness in different visual elements.

In the multi-candidate setting, which more closely mirrors real-world retrieval-augmented generation (RAG) systems, the 7B model achieves 32.3% bounding box accuracy and 41.8% answer accuracy when handling three candidate documents This demonstrates the model's capability to identify relevant sources among multiple documents while enabling fine-grained attribution. It should be noted that this setting is more challenging than the single-oracle candidate scenario, as visual source attribution among multiple candidates additionally requires the model to identify the relevant document among hard negatives.

When the oracle candidate is absent from the multi-candidate set, the model generally handles the "No Answer" scenario well. For example, VISA-7B-multi correctly indicates "No Answer" in 82.2% of cases on average for Wiki-VISA, refus-

Paradigm	Model	Paper-VISA Average bbx	ans	Passage bbx	ans	Non-Passage bbx	ans
Zero-Shot Prompt							
Textual	Qwen2-VL-72B	43.5	16.7	58.2	19.6	28.7	13.7
Visual	Qwen2-VL-72B	1.5	43.1	0.5	40.2	2.5	45.9
Fine-tune							
Textual	Qwen2-VL-2B	56.8	14.6	51.9	17.8	61.7	11.4
Textual	Qwen2-VL-7B	59.5	18.1	52.4	21.3	66.5	14.8
Visual	Qwen2-VL-2B	63.0	38.3	50.6	34.4	75.3	42.1
Visual	Qwen2-VL-7B	68.2	43.8	58.1	41.6	78.2	45.9

Table 3: Comparing with Text-based Visual Attribution in single oracle candidate setting of Paper-VISA. The visual paradigm indicates our VISA method. The textual method combines layout detector, OCR, and LLM.

Train Data	Wiki-VISA Average bbx	ans	Paper-VISA Average bbx	ans
Wiki	54.2	65.2	27.8	36.2
Paper	0.2	42.6	68.2	43.8
FineWeb	37.6	50.2	22.0	43.3
Wiki+Fineweb	58.2	65.3	21.0	43.1
Paper+Fineweb	36.1	48.7	66.5	44.6
Wiki+Paper+Fineweb	58.1	64.8	67.6	44.3

Table 4: Effectiveness of VISA trained on different combinations of training data for bounding box accuracy (bbx) and answer accuracy (ans) in the single oracle candidate setting.

ing to respond. In these cases, both attribution and answer are considered correct. In the remaining cases, the model attempts to answer despite lacking oracle evidence, leading to incorrect bounding boxes. Notably, in 8.8% of cases (91% - 82.2%), the model provides correct answers despite no oracle document in candidates, likely due to memorization, hallucination, or false negatives in candidate documents. This phenomenon does not occur in Paper-VISA, likely because synthetic queries in the publication domain are more directly related to the oracle document, whereas NaturalQuestions for Wiki-VISA are more general.

6 Analysis

6.1 Text-based Visual Source Attribution

Our VISA method performs RAG and visual source attribution in an end-to-end manner. Alternatively, a modularized text-based RAG pipeline—incorporating a layout prediction model, OCR, and text-based LLMs—could achieve similar functionality. Comparing VISA with such a modularized text-based pipeline would be valuable for understanding the advantages of different approaches. We construct a text-based pipeline for evaluating Paper-VISA in a single-oracle candidate setting. Using PubLayNet's bounding boxes, we assume a perfect layout model that accurately detects document elements. We apply `pytesseract` OCR to extract text from each bounding box, then feed the text list and a given question into an LLM, which generates an answer along with the index of the supporting evidence. The corresponding bounding box is used for visual attribution. For a fair comparison, we use Qwen2-VL's language model for both zero-shot prompting and fine-tuning.

As shown in Table 3, in the zero-shot setting, the text-based method achieves higher bounding box accuracy than the visual-based method, as LLMs are well-trained for text-based tasks. In contrast, visual-based methods struggle with precise bounding box attribution without fine-tuning. However, the text-based method has lower answer accuracy due to OCR errors, which introduce typos (e.g., misrecognized technical terms) making it more difficult to match the ground truth answer. For fine-tuned variants, both bounding box accuracy and answer accuracy improve over the zero-shot setting. However, the text-based method still has inherent limitations. While integrating a vision-language model (VLM) for OCR could potentially enhance text extraction accuracy, it introduces additional latency and complexity in the system. Moreover, in this comparison, we assume the text-based method benefits from a perfect layout detector—an unrealistic assumption in real-world applications. These findings further support the advantages of the proposed visual-based solution for VISA.

6.2 Out-of-Domain Zero-Shot

Table 4 shows the effectiveness of VISA while trained with different data combinations in the single candidate setting. It enables us to study the effectiveness of out-of-domain transfer and augmentation. First, we highlight the challenges of zero-shot generalization in VISA. Training and evaluating on in-domain achieves an effective bounding box accuracy, e.g. 54.2% on average for Wiki-VISA. However, significant performance drops are observed when models are tested on out-of-domain datasets. For instance, a model trained on Wiki-VISA achieves only 27.8% bounding box accuracy on Paper-VISA, while a model trained on Paper-VISA achieves near-zero performance (0.2%) on Wiki-VISA. This gap underscores the difficulty of transferring visual source attribution capabilities across datasets with differing document structures, layouts, and content modalities. Interestingly, Wiki-VISA appears to transfer better to Paper-VISA compared to the reverse. This may be because of the

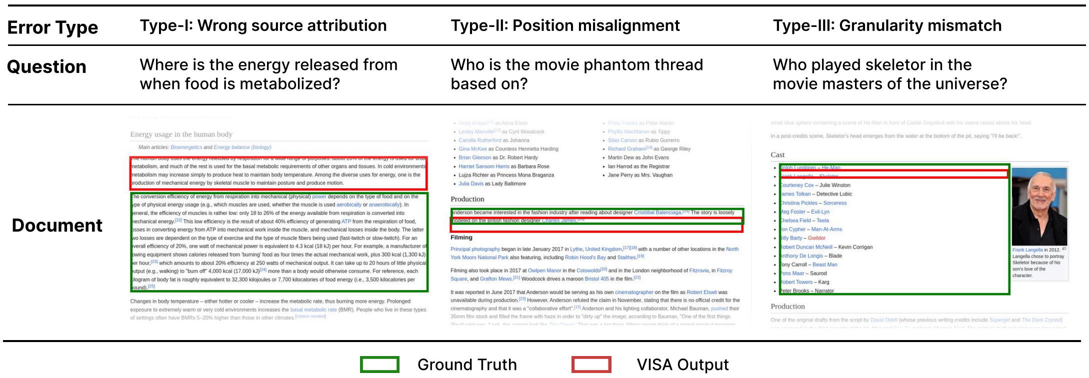


Figure 2: Type of errors in the evaluation of Wiki-VISA.

multi-page nature of Wiki-VISA, which provides richer training signals that generalize better to simpler single-page setting in Paper-VISA.

FineWeb-VISA shows as a promising resource for training models with improved zero-shot capabilities. When trained on FineWeb-VISA alone, the model achieves 37.6% bounding box accuracy on Wiki-VISA and 22.0% on Paper-VISA. Notably, FineWeb-VISA outperforms Wiki-VISA training on [>1] passage bbx accuracy for Wiki-VISA (57.3% vs. 50.1%), suggesting its effectiveness in handling long and complex document structures. However, FineWeb-VISA does not perform as well on non-passage content, likely due to its training focus on passage-level targets.

6.3 Data Augmentation

The results also demonstrate the benefits of augmenting training data with FineWeb-VISA. On Wiki-VISA, combining Wiki and FineWeb training data improves bounding box accuracy from 54.2% to 58.2% and improves performance on [>1] passages from 50.1% to 61.7%, indicating that FineWeb complements Wiki by enhancing the model's ability to attribute evidence in multi-page contexts. For Paper-VISA, however, augmenting with FineWeb does not significantly improve in-domain performance. Training on Paper+FineWeb achieves a comparable bounding box accuracy to Paper alone, but it enhances zero-shot performance on Wiki-VISA (from 0.2% to 36.1%).

Training on the full combination of datasets (Wiki+Paper+FineWeb) yields strong results across both domains, with 58.1% bbx accuracy on Wiki-VISA and 67.6% on Paper-VISA. This shows the importance of diverse training data for building generalizable models capable of handling different document types, layouts, and evidence modalities. Future work should focus on expanding the dataset diversity to further improve generalization and enable robust visual source attribution for a wide range of document structures.

6.4 Error Analysis

We conducted an error analysis on 50 randomly sampled cases from Wiki-VISA to better understand the limitations of VISA. Errors were categorized into three main types as demonstrated in Figure 2. The first type, wrong source attribution, occurred in 43 cases where the model attributed the source to an incorrect section of the document, failing to identify the precise region containing the evidence. The second type, position misalignment, was observed in 4 cases where the model appeared to have the correct intent but drew the bounding box inaccurately, either slightly off position or incorrectly sized. The third type, granularity mismatch, appeared in 3 cases where the model's attributed source, such as a specific cell in a table or an item in a list, did not match the ground truth granularity. While these cases could potentially be considered false negatives, we leave it in error analysis to emphasize the challenge in real-world use cases where user preferences for granularity may differ from the model's output.

7 Conclusion

We introduced VISA, a visual source attribution approach — generating answers while providing bounding boxes to locate evidence — for retrieval-augmented generation. Our curated datasets demonstrate its effectiveness across diverse document types, including complex multi-page and multimodal content. Experimental results show VISA bridges information retrieval and answer generation with finer-grained, visually grounded attribution. We hope VISA represents a pioneering step for more verifiable and user-friendly RAG systems.

8 Limitations

While VISA demonstrates promising results for answer generation and content grounding in vision-based RAG systems, it has several limitations.

Gap between our settings and real-world scenarios. Our approach focuses on generating short answers, which may not suffice for scenarios requiring detailed or explanatory responses, highlighting the need for enhancements in generating richer context. Besides, our curated datasets assume that answers are derived from a single, localized region within a document. However, in real-world applications, supporting evidence may span multiple sections or even multiple documents, limiting the model's effectiveness in more complex retrieval scenarios. Additionally, in the Natural Questions dataset (converted to our Wiki-VISA), short answers are often extracted substring from the evidence section. This presents another gap, as real-world answers may be implied by the evidence rather than being an exact substring.

Cross-domain generalization. Although our evaluation spans web and medical scientific papers containing diverse content modalities (e.g., passages, tables, and figures), it does not fully capture the variability of real-world documents, such as scanned or handwritten content. These often feature more complex layouts and diverse aspect ratios, posing additional challenges. Our zero-shot evaluation shows that while the model achieves reasonable bounding box accuracy in cross-domain transfer, its performance still lags behind in-domain effectiveness. Enhancing cross-domain generalization would make the VISA pipeline more robust for vision-based RAG tasks across a broader range of document types.

Trade-off between accuracy and efficiency. To create challenging attribution tasks, we designed Wiki-VISA images to contain content from four pages. However, increasing the candidate set further raises training costs and frequently leads to out-of-memory (OOM) issues given our limited computing resources. We hence the number of document candidates to three in our multi-candidate setting following previous practice (Yu et al., 2024). Our findings show a clear performance difference between single-image and multi-candidate settings, underscoring the challenge of scaling candidate size. In practical applications where VISA is integrated with retriever, further research is needed to balance candidate size, accuracy, and computational efficiency.

Aligning with real user expectation on visual attribution As briefly discussed in Section 6.4, a potential challenge lies in whether the visual attribution provided by VISA aligns with users' expectations in terms of granularity. Since VISA is designed to make answer verification more intuitive, conducting user studies in real-world deployment scenarios would provide deeper insights into its practical utility and potential refinements.

References

Akari Asai, Zexuan Zhong, Danqi Chen, Pang Wei Koh, Luke Zettlemoyer, Hannaneh Hajishirzi, and Wen tau Yih. 2024. Reliable, adaptable, and attributable language models with retrieval. *Preprint*, arXiv:2403.03187.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed question answering: Evaluation and modeling for attributed large language models. *Preprint*, arXiv:2212.08037.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020*, pages 213–229, Cham. Springer International Publishing.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Kanzhi Cheng, Qiushi Sun, Yougang Chu, Fangzhi Xu, Li YanTao, Jianbing Zhang, and Zhiyong Wu. 2024. SeeClick: Harnessing GUI grounding for advanced visual GUI agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9313–9332, Bangkok, Thailand. Association for Computational Linguistics.

Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. 2024. M3DocRAG: Multi-modal retrieval is what you need for multi-page multi-document understanding.

Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. 2016. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

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

Jeremy J. Foster. 1979. *The Use of Visual Cues in Text*, pages 189–203. Springer US, Boston, MA.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv:2312.10997*.

Muhammad Khalifa, David Wadden, Emma Strubell, Honglak Lee, Lu Wang, Iz Beltagy, and Hao Peng. 2024. Source-aware training enables knowledge attribution in language models. In *First Conference on Language Modeling*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Kevin Qinghong Lin, Linjie Li, Difei Gao, Zhengyuan Yang, Zechen Bai, Weixian Lei, Lijuan Wang, and Mike Zheng Shou. 2024. ShowUI: One vision-language-action model for generalist GUI. In *NeurIPS 2024 Workshop on Open-World Agents*.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA. Association for Computational Linguistics.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2024. GPT-4o system card. *arXiv:2410.21276*.

Oded Ovadia, Menachem Brief, Moshik Mishaeli, and Oren Elisha. 2024. Fine-tuning or retrieval? comparing knowledge injection in LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 237–250, Miami, Florida, USA. Association for Computational Linguistics.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv:2406.17557*.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. 2024. Glamm: Pixel grounding large multimodal model. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13009–13018.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Monica Riedler and Stefan Langer. 2024. Beyond text: Optimizing rag with multimodal inputs for industrial applications. *Preprint*, arXiv:2410.21943.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv:2412.03555*.

John Sweller. 2011. Chapter two - cognitive load theory. volume 55 of *Psychology of Learning and Motivation*, pages 37–76. Academic Press.

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. *arXiv:2409.12191*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *Preprint*, arXiv:2410.13085.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *Preprint*, arXiv:2410.10594.

Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. 2023. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv:2307.08581*.

Zhong-Qiu Zhao, Peng Zheng, Shou tao Xu, and Xindong Wu. 2019. Object detection with deep learning: A review. *Preprint*, arXiv:1807.05511.

Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. 2019. PubLayNet: Largest Dataset Ever for Document Layout Analysis . In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022, Los Alamitos, CA, USA. IEEE Computer Society.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. 2023. Object detection in 20 years: A survey. *Preprint*, arXiv:1905.05055.

Train Data	Wiki-VISA		Paper-VISA	
	bbx	ans	bbx	ans
Crop, Absolute	54.2	65.2	27.8	36.2
No Random Crop	58.8	65.6	1.7	36.9
Normalized Value	56.4	64.4	0.1	37.2
No Bounding Box	0	67.6	0	35.2

Table 5: Impact of bounding box target representation and cropping strategies during training on Wiki-VISA in the single oracle candidate setting.

Model	Avg		[<1] Passage		[>1] Passage		non-Passage	
	bbx	ans	bbx	ans	bbx	ans	bbx	ans
Zeroshot Prompt								
Qwen2-72B-VL	1.5	60.4	3.4	58.5	0.1	54.9	0.9	67.9
gpt4o	0.0	52.8	0.0	50.9	0.0	43.3	0.0	64.3
Fine-tune								
QWen2-VL-2B	37.5	57.1	70.0	61.1	18.7	44.9	23.8	65.3
QWen2-VL-7B	54.2	65.2	75.6	66.5	50.1	56.0	36.8	73.1
Phi3-Vision	34.0	49.8	59.9	54.5	19.1	40.2	22.9	54.6

Table 6: Effectiveness of VISA prompted or finetuned with model other than Qwen2-VL in the single oracle candidate setting on Wiki-VISA.

A Appendix

A.1 Dataset Licenses

- NQ: Apache License 2.0
- Wikipedia: Creative Commons Attribution Share Alike, GNU Free Documentation License family.
- Fineweb-edu: Open Data Commons License Attribution family.
- PubLayNet: Community Data License Agreement – Permissive, Version 1.0.
- VISA Datasets: Our crafted datasets follow the same license as the source of the documents.

A.2 Model Backbone Licenses

- Qwen2-VL-72B-Instruct: Qwen LICENSE AGREEMENT.
- Qwen2-VL-2B-Instruct: Apache License.
- Qwen2-VL-7B-Instruct: Apache License.
- VISA Models: Our fine-tuned models follow the same licenses as the original model backbone.

A.3 Bounding Box Target

Table 5 shows the impact of different bounding box target representations and cropping strategies during training. Training with random cropping and absolute coordinate values achieves a balance between in-domain performance on Wiki-VISA (54.2%) and zero-shot generalization to Paper-VISA (27.8%) in bounding box accuracy. Removing random cropping slightly improves Wiki performance but drastically reduces zero-shot generalization, indicating that random cropping enhances the model's robustness to varied input sizes. Normalizing coordinate values achieves moderate performance on Wiki-VISA but fails on Paper-VISA, suggesting that absolute bounding box values are better suited to our experiments.

The "No Bounding Box" row represents a vanilla visual retrieval-augmented generation setup without visual source attribution, where models generate answers without bounding box predictions. VISA enables visual source attribution capability while the effectiveness of answer generation is preserved at about the same level of effectiveness.

A.4 Training Hyper-parameters

The training objective for both single-candidate and multi-candidate setting are next-token prediction with cross-entropy loss. We fine-tuned the models for two epochs in the single-candidate setting, using LoRA with a learning rate of 1e-4, a batch size of 64, and 4×H100 GPUs. For the multi-candidate setting, we initialized the models with weights from the single-candidate setup and trained for one epoch with the same learning rate. We froze the image encoder to reduce GPU memory usage in the multi-candidate setting.

A.5 Model Backbone Choice

Beyond the QWen2-VL-Instruct series, we also explored prompt GPT4o for zero-shot visual source attribution or fine-tuning Phi3-Vision-Instruct on the single oracle candidate setting on the Wiki-VISA dataset. As shown in 6, the QWen2-VL-Instruct series performs better in the VISA task.

A.6 Detailed Results of Data Effectiveness

We provide the detailed results of Table 4 in Table 7.

A.7 Zero-Shot Prompting

In addition to the zero-shot prompt 72B Qwen2-VL-Instruct model, as in Table 2, we further explored zero-shot prompt 2B and 7B variants of Qwen2-VL-Instruct model as shown in Table 8. These results indicate similar trends as seen with

Train Data	Wiki-VISA								Paper-VISA					
	Average		[<1] Passage		[>1] Passage		Non-Passage		Average		Passage		Non-Passage	
	bbx	ans	bbx	ans	bbx	ans	bbx	ans	bbx	ans	bbx	ans	bbx	ans
Wiki	54.2	65.2	75.6	66.5	50.1	56.0	36.8	73.1	27.8	36.2	20.5	32.6	35.1	39.7
Paper	0.2	42.6	0	46.3	0.4	33.5	0.1	48.1	68.2	43.8	58.1	41.6	78.2	45.9
FineWeb	37.6	50.2	48.9	45.1	57.3	52.3	6.6	53.1	22.0	43.3	26.5	41.7	17.4	44.9
Wiki+Fineweb	58.2	65.3	68.7	66.6	61.7	57.1	44.1	72.1	21.0	43.1	18.5	42.2	23.4	43.9
Paper+Fineweb	36.1	48.7	51.8	49.6	49.6	44.2	6.8	52.4	66.5	44.6	56.1	42.2	76.9	47.0
Wiki+Paper+Fineweb	58.1	64.8	69.9	65.0	58.7	56.7	45.8	72.7	67.6	44.3	55.9	41.5	79.3	47.1

Table 7: Effectiveness of VISA trained on different combinations training data for bounding box accuracy (bbx) and answer accuracy (ans) in the single oracle candidate setting.

Single Oracle Candidate	[<1] Passage		[>1] Passage		non-Passage	
	bbx	ans	bbx	ans	bbx	ans
Qwen2-2B-VL (zeroshot prompt)	0.1	30.7	0.0	22.7	0.0	35.9
Qwen2-7B-VL (zeroshot prompt)	1.7	52.0	0.1	39.7	0.1	57.8

Table 8: Effectiveness of prompting Qwen2-VL-2B and 7B in zero-shot, in the single oracle candidate setting in Wiki-VISA.

larger models: zero-shot prompting methods are not ready to effectively conduct the VISA task.

A.8 Prompt for Single Oracle candidate VISA

The following prompt template was used to format the model's inputs and outputs for training the *Single Oracle Candidate* VISA.

```
Model Input:
System:
Given a document image, your task is to answer the question and locate the source of the answer via a bounding box.

User:
{image} Image Size: {image.size}
Question: {question}

Model Output:
Assistant:
Answer: {answer}
Bounding Box: {bounding_box}
```

We also explored different prompting strategies that swap the order of Answer and Bounding Box in the above prompt template and the comparison is shown in Table 9.

Single Oracle Candidate	[<1] Passage		[>1] Passage		non-Passage	
	bbx	ans	bbx	ans	bbx	ans
VISA-7B-Single	75.6	66.5	50.1	56.0	36.8	73.1
VISA-7B-Single-Swap	72.8	65.0	44.0	53.9	34.8	69.3

Table 9: Comparision between using the above prompt template (VISA-7B-Single) and swapping the order of Answer and Bounding Box. (VISA-7B-Single-Swap)

A.9 Prompt for Multi-candidate VISA

The following prompt template was used to format the model's inputs and outputs for training the *Multi-candidate* VISA.

```
Model Input:
System:
Given document images, your task is to answer the question and locate the source of the answer via a bounding box.

User:
{image1} Image Size: {image1.size}
{image2} Image Size: {image2.size}
{image3} Image Size: {image3.size}
Question: {question}

Model Output:
Assistant:
Answer: {answer}
Evidence Document: {index}
Bounding Box: {bounding_box}
```

A.10 Prompt for synthetic data generation

The following prompt was used for prompting QWen2-VL-72B to generate synthetic questions and answers for Paper-VISA and Fineweb-VISA datasets.

```
System:
Ask a question that can be specifically answered by the content in the red bounding box area and give a short answer. The question can be a wh- question, a yes/no question, or a how question, that can be answered in a few words.
Output format:

Question: <question>
Short Answer: <short answer>

Or simply return 'Empty' if the bounding box area is not visible or informative.

User: {image}
```

An example of synthetic data from Paper-VISA can be found in Figure 3.

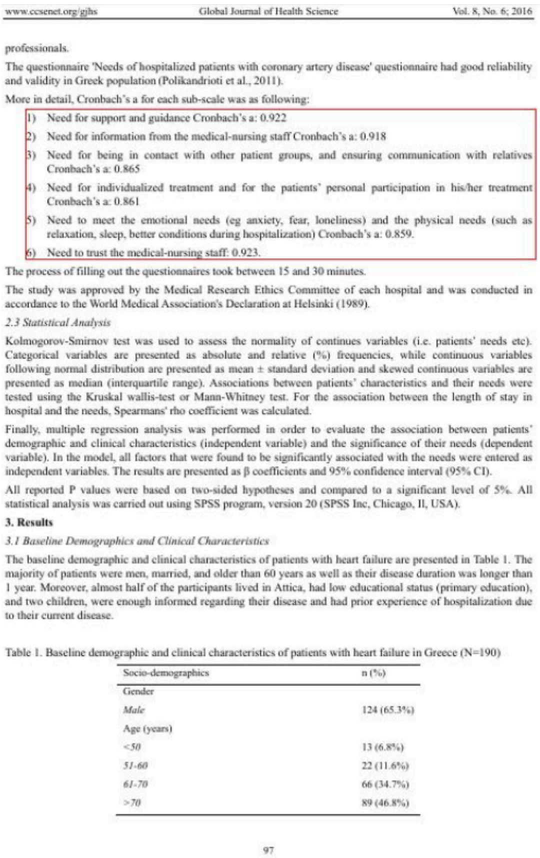
A.11 AI Assistant Usage

GPT4o is used during the writing to correct grammar errors and format tables.

A.12 Additional Qualitative Examples

In Figure 4 and Figure 5, we provide additional qualitative examples of the success and failure cases of our model VISA.

Input document screenshot with bounding box



www.ccsenet.org/gjhs Global Journal of Health Science Vol. 8, No. 6; 2016

professionals.

The questionnaire 'Needs of hospitalized patients with coronary artery disease' questionnaire had good reliability and validity in Greek population (Polikandrioti et al., 2011).

More in detail, Cronbach's a for each sub-scale was as following:

1) Need for support and guidance Cronbach's a: 0.922
2) Need for information from the medical-nursing staff Cronbach's a: 0.918
3) Need for being in contact with other patient groups, and ensuring communication with relatives Cronbach's a: 0.865
4) Need for individualized treatment and for the patients' personal participation in his/her treatment Cronbach's a: 0.861
5) Need to meet the emotional needs (eg anxiety, fear, loneliness) and the physical needs (such as relaxation, sleep, better conditions during hospitalization) Cronbach's a: 0.859.
6) Need to trust the medical-nursing staff: 0.923.

The process of filling out the questionnaires took between 15 and 30 minutes.

The study was approved by the Medical Research Ethics Committee of each hospital and was conducted in accordance to the World Medical Association's Declaration at Helsinki (1989).

2.3 Statistical Analysis

Kolmogorov-Smirnov test was used to assess the normality of continues variables (i.e. patients' needs etc). Categorical variables are presented as absolute and relative (%) frequencies, while continuous variables following normal distribution are presented as mean ± standard deviation and skewed continuous variables are presented as median (interquartile range). Associations between patients' characteristics and their needs were tested using the Kruskal wallis-test or Mann-Whitney test. For the association between the length of stay in hospital and the needs, Spearmans' rho coefficient was calculated.

Finally, multiple regression analysis was performed in order to evaluate the association between patients' demographic and clinical characteristics (independent variable) and the significance of their needs (dependent variable). In the model, all factors that were found to be significantly associated with the needs were entered as independent variables. The results are presented as β coefficients and 95% confidence interval (95% CI).

All reported P values were based on two-sided hypotheses and compared to a significant level of 5%. All statistical analysis was carried out using SPSS program, version 20 (SPSS Inc, Chicago, Il, USA).

3. Results

3.1 Baseline Demographics and Clinical Characteristics

The baseline demographic and clinical characteristics of patients with heart failure are presented in Table 1. The majority of patients were men, married, and older than 60 years as well as their disease duration was longer than 1 year. Moreover, almost half of the participants lived in Attica, had low educational status (primary education), and two children, were enough informed regarding their disease and had prior experience of hospitalization due to their current disease.

Table 1. Baseline demographic and clinical characteristics of patients with heart failure in Greece (N=190)

Socio-demographics	n (%)
Gender	
Male	124 (65.3%)
Age (years)	
<50	13 (6.8%)
51-60	22 (11.6%)
61-70	66 (34.7%)
>70	89 (46.8%)

97

Generated question and answer

Question: What is the Cronbach's alpha for the need for support and guidance sub-scale?

Short Answer: 0.922

Figure 3: An example of synthetic data from Paper-VISA.

Figure 4: Success examples of visual source attribution on Wiki-VISA by VISA-7B-single.

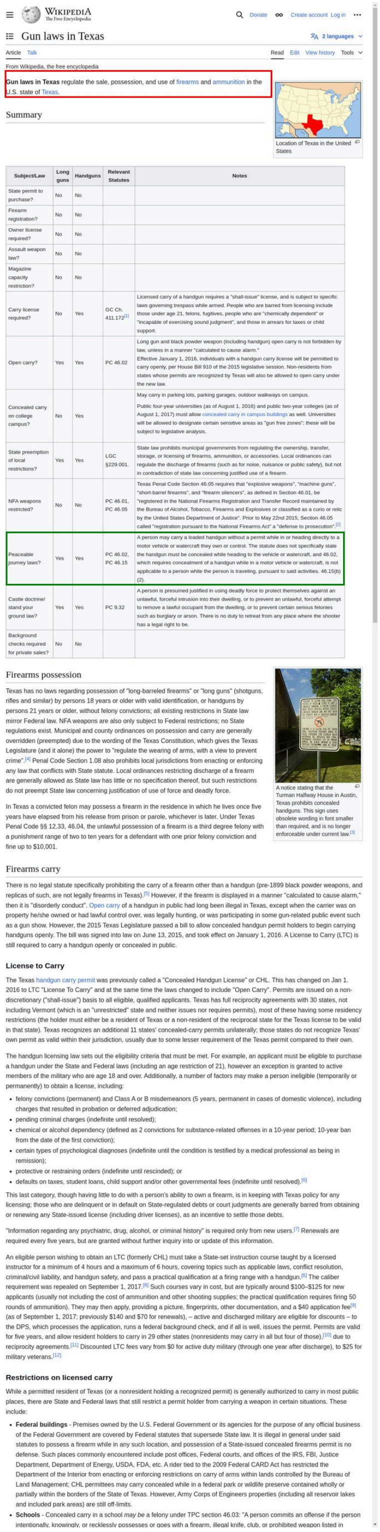
Failure Cases

Ground Truth   VISA Output

Q: Is it legal to carry a gun in a car in Texas?

Q: How much does the second place winner get in the US Open?

Q: how many episodes in season 2 of the durrells in corfu?



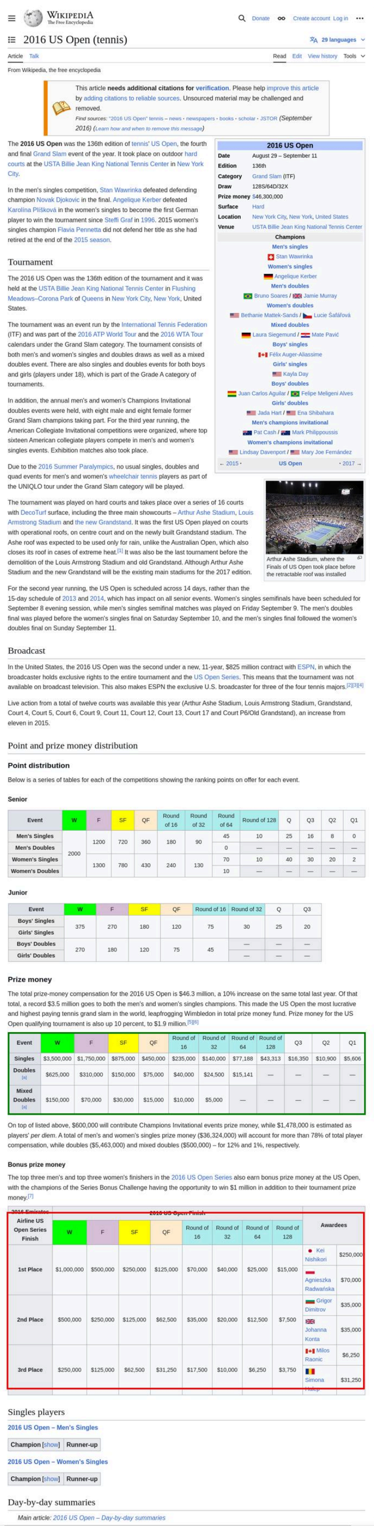


Figure 5: Failure examples of visual source attribution on Wiki-VISA by VISA-7B-single.