Look As You Think: Unifying Reasoning and Visual Evidence Attribution for Verifiable Document RAG via Reinforcement Learning

Supplementary Material

Visual evidence attribution in visual document retrieval-augmented generation (VD-RAG) requires models to both generate accurate answers and identify supporting evidence within external visual documents.

We propose the LAT framework, which enables Chain-of-Evidence (CoE) reasoning through a two-stage training paradigm. In Stage I, the model is fine-tuned on a set of human-verified CoE annotations to acquire effective reasoning patterns. Next, it undergoes reinforcement learning (RL) with tailored reward functions that guide stepwise attribution. This design enhances both reasoning capability and attribution quality, thereby achieving robust stepwise attribution and consistent performance across single- and multi-image scenarios without relying on process-level ground-truth annotations.

A Experiment Details

Since CoE reasoning requires flexible grounding capabilities and adaptability in low-resource conditions, we conducted preliminary experiments to identify a suitable backbone model. Specifically, we evaluated Qwen2.5-VL and InternVL2.5 by prompting them to answer questions with direct evidence attribution. As shown in Table 8, Qwen2.5-VL achieved a 2.1% improvement in EM under zero-shot prompting, whereas InternVL2.5 exhibited performance degradation. The weaker performance of InternVL2.5 may stem from its reliance on special tokens (e.g., `<box>`, `<ref>`) for grounding, which constrains the flexible grounding necessary for CoE reasoning in zero-shot settings and impedes effective convergence when training under low-resource conditions.

In contrast, Qwen2.5-VL represents bounding boxes in JSON format, allowing more flexible and direct visual evidence attribution without the constraints of special tokens. This design aligns well with the requirements of CoE reasoning. Building on this foundation, we introduce the two-stage training framework to enhance evidence grounding and reasoning consistency.

Stage I Cold Start: We sampled instances from each dataset (Paper-VISA, Wiki-VISA, FineWeb-VISA) and filtered them using the recall metric defined in Section 3 to retain correct samples. Manual correction was then performed to address bounding box drift by adjusting their positions and sizes to ensure alignment with the correct content regions. The final dataset splits are summarized in Table 6 (cf. Table 5 for data usage comparison). We employed LoRA for parameter-efficient fine-tuning and maintained these configurations throughout training.

Stage II RL Training: We employed GRPO as the policy optimization framework, with the reward function defined in Section 3. To ensure training stability, the learning rate was set to $5\mathrm{e}{-5}$ for both Paper- and Wiki-VISA, and training was conducted for one epoch. We further configured 8 and 6 rollouts for Paper- and Wiki-VISA, respectively, to enhance sampling diversity. The maximum prompt and completion lengths were limited to 16,384 and 600 tokens.

For the stepwise attribution reward R_{step}, we required $R_{\text{acc}} \geq \epsilon$, which zeroes the step rewards for incorrect answers and thereby reduces the advantage of erroneous samples. As specified in Equation 3, if the predicted answer a exactly matches the ground truth a_{gt} (EM = 1), the recall is 1 and the reward is set to 1. For non-exact matches, recall was computed as the proportion of overlapping tokens between a and a_{gt}, normalized by the ground truth length (Equation 2). Manual evaluation of a representative subset indicates that answers require at least 80% token overlap to preserve semantic completeness ($\gamma = 0.8$). This corresponds to the accuracy reward R_{acc} of approximately 0.4 ($\frac{\mathbb{I}(\text{EM}(a,a_{gt})=1)+\text{Recall}(a,a_{gt})}{2} = \frac{0+0.8}{2}$). Accordingly, we set $\epsilon = 0.4$ as the threshold, ensuring that faithful reasoning towards the answer is rewarded.

To promote attribution diversity across reasoning steps, we applied IoU-based constraints with a threshold $\gamma = 0.5$. A predicted box was regarded as valid evidence if its IoU with the corresponding ground truth exceeded 0.5, indicating sufficient relevance. Conversely, pairwise IoU among predicted boxes was required to remain below 0.5, ensuring that distinct reasoning steps relied on independent, non-overlapping evidence regions.

Implementation Details: All experiments are conducted using Qwen2.5-VL-7B-Instruct on 4×A800 (80 GB) GPUs,

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

Table 5: Statistics for the entire VISA dataset, where the construction quantity for multi-image settings is consistent with that for single-image settings.

Dataset	# Train-LAT		# Test
	Cold-Start Stage	RL Stage	
Single-Image / Multi-Image			
Wiki-VISA	264 / 463	4,676 / 4,676	3,000
Paper-VISA	271 / 355	5,000 / 3,562	2,160
Fineweb-VISA	108 / 124	2,000 / 2,000	-

Table 6: Training subsets sampled from the Paper- and Wiki-VISA datasets under single- and multi-image settings.

Stage	Parameter	Value
Cold Start	γ	0.8
	learning rate	1e−4
	batch Size	8
	epoch(s)	2
	lora rank (r)	64
	lora scaling (α)	64
	lora dropout	0.05
RL Training	learning rate	5e−5
	batch size	16 / 8
	epoch(s)	1
	rollout	8 / 6
	τ	0.3
	δ	0.5
	ϵ	0.4
	lora rank (r)	64
	lora scaling (α)	64
	lora dropout	0.05

Table 7: Hyperparameter Settings for Training. Values marked with A / B correspond to different configurations used for Paper- and Wiki-VISA, respectively.

requiring approximately 20 GPU hours for single-domain training and 48 GPU hours for multi-domain settings. Training leverages mixed precision with DeepSpeed (Rasley et al. 2020; Rajbhandari et al. 2021) and CPU offloading for memory efficiency. To mitigate computational overhead, we employed Flash-attention2 (Dao 2023) and gradient checkpointing techniques. Detailed hyperparameters for different training stages are presented in Table 7. The pseudocode for LAT is provided in Algorithm 1. For multi-image scenarios, initialize the model from the single-image RL-trained parameters, then further apply SFT on multi-image CoE data in $\mathcal{D}_{\text{final}}$ following the above procedures.

For reproducibility, we fixed the random seed to 3407 during training and disabled sampling at evaluation by setting `do_sample=False`. In contrast to VISA (Ma et al. 2024b), we maintained original image resolutions since our CoE reasoning framework requires comprehensive visual understanding across the entire image.

Algorithm 1: LAT Framework

Input: Sampled query set $\mathcal{Q}_{\text{cold\_start}}$ and $\mathcal{Q}_{\text{RL}}$
Parameter: Learning rates η_1, η_2; Reward thresholds $\tau, \delta, \epsilon, \gamma$
Output: LAT model π_θ

Stage I: Cold-start Supervised Fine-tuning
for each query $q \in \mathcal{Q}_{\text{cold\_start}}$ **do**
  Generate CoE reasoning traces using Gemini2.5 pro
  Filter outputs by the recall metric
  if Recall$(a, a_{gt}) \geq \gamma$ **then**
    $\mathcal{D}_{\text{final}} \leftarrow \mathcal{D}_{\text{final}} \cup \{q\}$
  end if
end for
Manually verify bounding boxes and format
Train model π_θ via SFT on verified data $\mathcal{D}_{\text{final}}$ with learning rate η_1
Stage II: Reinforcement Learning with GRPO
for each training step **do**
  Sample a batch of queries $q \sim \mathcal{Q}_{\text{RL}}$
  Model π_θ generates CoE reasoning steps $\{r_1, ..., r_T\}$ and bounding boxes $\{B_1, ..., B_T\}$ for each query q
  Compute reward $R = R_{\text{acc}} + R_{\text{step}} + R_{\text{ground}} + R_{\text{format}}$
  Update policy π_θ using GRPO with R and learning rate η_2
end for
return final LAT model π_θ

B Dataset Construction

VISA Dateset

1) Wiki-VISA: Selenium renders Wikipedia pages for Natural Questions (NQ) (Kwiatkowski et al. 2019) query-answer pairs, with HTML elements (containing answers) and their bounding boxes annotated.

2) Paper-VISA: Based on PubLayNet (Zhong, Tang, and Yepes 2019). Vision-language models (VLMs) generate QA pairs from layout-annotated scientific documents, with answer-region bounding boxes extracted.

3) FineWeb-VISA: Sampled from FineWeb-edu (Penedo et al. 2024). VLMs generate queries/short answers for educational webpage passages longer than 50 tokens, with screenshots and answer-region bounding boxes.

Multi-Image Data Construction

To simulate real-world VD-RAG scenarios, VISA constructs a multi-image document experimental environment after obtaining the query-document-answer-bounding box triplets. Specifically, given a query q, VISA employs a retriever to retrieve top-k candidate documents, then randomly samples $m-1$ hard negative candidates that do not contain the ground truth. These negative samples are combined with one source document containing the correct answer to serve as input for the multi-image scenarios. VISA deliberately avoids directly utilizing the top-m retrieval results to prevent bias toward specific retrievers or candidate document positions, thereby ensuring methodological generalizability.

To evaluate the model's capability in handling no-answer scenarios, VISA randomly replaces the source document in the candidate set with a 20% probability, simulating realistic cases where the retriever fails to return documents containing the correct answer. In the specific experimental setup, VISA leverages the Document Screenshot Embedding (DSE) model as the retriever. The parameters are set to $k = 20$ and $m = 3$. Example cases are shown in Figure 5.

Training Data Preprocess

We employed Gemini2.5 Pro (Comanici et al. 2025) to generate CoE data. We used the prompt shown in Figure 9 to guide stepwise evidence attribution during generation.

We subsequently applied a recall metric to measure response against reference answers for preliminary sample filtering. To ensure data quality, we manually reviewed and corrected instances of irrelevant attributions, spatial coordinate offsets, and format inconsistencies. The resulting dataset $\mathcal{D}_{\text{final}}$ is then employed for cold-start training. Examples are illustrated in Figure 6 and 7.

Models	Param.	Wiki-VISA (Single)		Paper-VISA (Single)		Wiki-VISA (Multi)		Paper-VISA (Multi)	
	Size	EM	IoU@0.5	EM	IoU@0.5	EM	IoU@0.5	EM	IoU@0.5
Open-source Models, Direct Answer									
InternVL2.5	8B	45.9	-	42.4	-	37.7	-	32.1	-
Qwen2.5-VL	7B	67.7	-	38.2	-	54.1	-	34.6	-
Open-source Models, Direct Answer & Attribution (DA)									
InternVL2.5	8B	45.1	0	42.0	0.60	35.0	1.67	31.4	1.94
Qwen2.5-VL	7B	69.4	0.80	43.8	2.87	52.5	0.97	37.2	5.56

Table 8: Performance comparison of various models for direct answer generation and attribution tasks. "DA" refers to the zero-shot prompting setting for direct answer & attribution without training.

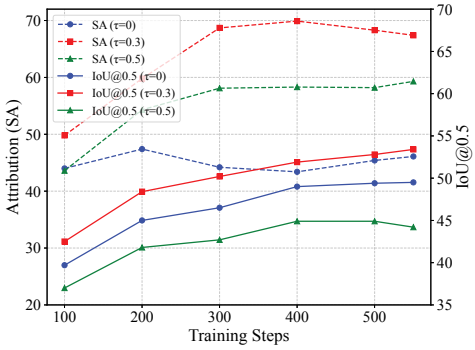


Figure 4: Performance variation with different τ (Wiki).

C Prompt

We designed task-specific prompts adapted to reasoning strategies and model architectures. For models such as LLaVA-CoT (Xu et al. 2025) and R1-OneVision (Yang et al. 2025), which are trained to take questions directly as input, with no additional prompting. Their performance is evaluated by extracting content from the generated outputs, specifically from segments enclosed within the `<CONCLUSION>` and `<answer>` tags. Since both Gemini and InternVL generate bounding boxes in normalized coordinates. During evaluation, these normalized values are converted to absolute coordinates based on the size of the source images. For other models, we employed prompts that initiate responses with "The answer is:", which encourages concise answers without irrelevant explanations.

Each bounding box must be paired with its source image index using the JSON format `{"bbox_2d": [x_1, y_1, x_2, y_2], "image_index": i}`. As illustrated in Figure 10, we employed a consistent prompt format across both the cold-start and RL training stages.

D Baseline

Our results demonstrate that LAT consistently outperforms open-source models of comparable or larger scales, including InternVL2.5-8B (Chen et al. 2025c), Qwen2.5-VL-7/32B-Instruct (Bai et al. 2025a), LLaVA-OneVision (Li et al. 2024), mPLUG-DocOwl2 (Hu et al. 2024), LLaVA-CoT (Xu et al. 2025), and R1-OneVision (Yang et al. 2025). We note that LLaVA-CoT is trained exclusively on single-image datasets without multi-image reasoning supervision and is therefore excluded from the multi-image evaluation.

LAT achieves superior performance compared to certain closed-source models, including Qwen-VL Max (Bai et al. 2023) and Gemini 2.0 Flash (Anil et al. 2025). In our experimental setting, VISA-7B is performed with the open-source implementation available on HuggingFace[1]. GCoT(Wu et al. 2025) and VisCoT(Shao et al. 2024a) rely on single-image datasets with costly step annotations (e.g., 438k) and are not trained in multi-image settings. The EM of VisCoT on Paper/Wiki-VISA is 10.5/8.8%, illustrating the domain gap.

E Further Analysis

Sensitivity analysis of attribution threshold τ.

To ensure the alignment quality of visual evidence-text pairs in CoE reasoning, we set a threshold τ in the stepwise attribution recall reward function to distinguish positive from negative samples. Using the synthetic CoE data $\mathcal{D}_{\text{final}}$, we computed the values of semantic similarity on manually corrected annotations. As shown in Figure 8, we recorded the maximum and minimum similarities for samples grouped by answer type in each dataset, excluding pairs with zero similarity. Based on this analysis, we selected 0.3 as the threshold, slightly above the minimum similarity observed in human-corrected annotations.

To examine the effect of different thresholds, we conducted ablation experiments on single-image scenarios from Paper-VISA (Figure 3c) and Wiki-VISA (Figure 4). These experiments reveal that the choice of τ has an impact on reasoning outcomes. For example, a relatively high threshold, such as 0.5, leads to notable declines in both accuracy and IoU@0.5, and the overall process traceability quality (SA) also fails to improve, because such strict thresholds make it difficult to sample sufficient positive instances.

Generalization and Ablation

The results in Table 3 demonstrate LAT's cross-domain transferability, achieving superior performance over SFT with improvements of 1.7% in EM and 6.2% in IoU@0.5 when transferring from Paper- to Wiki-VISA datasets. In the reverse transfer (Wiki→Paper), a slight drop in IoU@0.5 is

[1]https://huggingface.co/collections/MrLight/visa-rag-with-visual-source-attribution

Train→Eval	Method	EM	IoU@0.5
Wiki→Paper (Single)	SFT	36.9	11.5
	LAT-Ind.	$43.6_{\uparrow 6.7}$	$10.4_{\downarrow 1.1}$
Wiki→Paper (Multi)	SFT	34.3	5.5
	LAT-Ind.	$44.3_{\uparrow 10.0}$	$21.9_{\uparrow 16.4}$

Table 9: LAT demonstrates robust generalization with cross-domain transfer between Paper-VISA and Wiki-VISA.

Dateset	Method	Total	Correct	Acc.
Paper	LAT-Ind.	425	356	0.84
	LAT-Full	425	343	0.81
Wiki	LAT-Ind.	578	267	0.46
	LAT-Full	578	283	0.49

Table 10: The accuracy of correctly detecting no-answer cases in multi-image settings.

observed. This can be attributed to the training bias on Wiki-VISA, which contains high-resolution images with relatively dispersed layouts, whereas Paper-VISA features compact medical documents. Such dense layouts often lead to incomplete evidence localization. Nevertheless, the model maintains strong reasoning consistency and answer accuracy, preserving the vanilla model's performance on both Wiki-VISA (EM: 67.7%) and Paper-VISA (EM: 38.2%). The generalization is even more apparent in multi-image scenarios (Table 3 and 9). These results indicate that LAT improves transfer effectiveness while retaining adaptability across diverse document types.

We also conducted further ablation studies. Without the cold-start stage, the performance achievable by RL is fundamentally capped (EM, 43.9%; IoU, 24.1%; SA, 33.4% on Paper-VISA). Moreover, the learning curves indicate that RL training for a single epoch has not yet reached its performance ceiling. We therefore extended the training to 2 epochs on Paper-VISA, achieving improved results (EM, $46.5_{\uparrow 1.1}\%$; IoU, $50.1_{\uparrow 0.2}\%$; SA, $38.1_{\uparrow 2.6}\%$).

F Case Study

Figure 12–17 present representative examples of model responses. Overall, LAT improves the traceability of the reasoning process. Based on the evaluation of several model responses, we find that LAT generates coherent Chain-of-Evidence (CoE) reasoning traces while maintaining general QA performance and attribution precision. The model is guided by the prompt and CoE training data to directly locate content relevant to the query, following a coarse-to-fine observation process as shown in Figures 12 and 13. In Wiki-VISA, where reasoning requires searching across different layouts, CoE enables the model to verify answer correctness by progressively narrowing down evidence. Figures 16 and 17 show the results under multi-image scenarios. Table 10 reports the accuracy of correctly detecting no-answer cases in multi-image settings.

G Limitation and Future Work

While LAT substantially improves stepwise attribution and overall performance, several practical aspects remain open for future enhancement. First, the thresholds used in the stepwise attribution rewards are manually set based on annotated examples, a choice that offers training stability. Exploring adaptive thresholding mechanisms, such as confidence-aware scaling, may improve flexibility and enhance alignment robustness. Second, LAT is currently trained only on the VISA dataset, whose evidence annotations predominantly follow a single-hop structure. As a result, LAT's ability to generalize to multi-hop or cross-source evidence reasoning is not yet fully evaluated or supervised. Incorporating datasets with explicit multi-hop evidence chains would provide the necessary supervision to assess and strengthen LAT's capacity for longer reasoning trajectories and cross-page evidence composition.

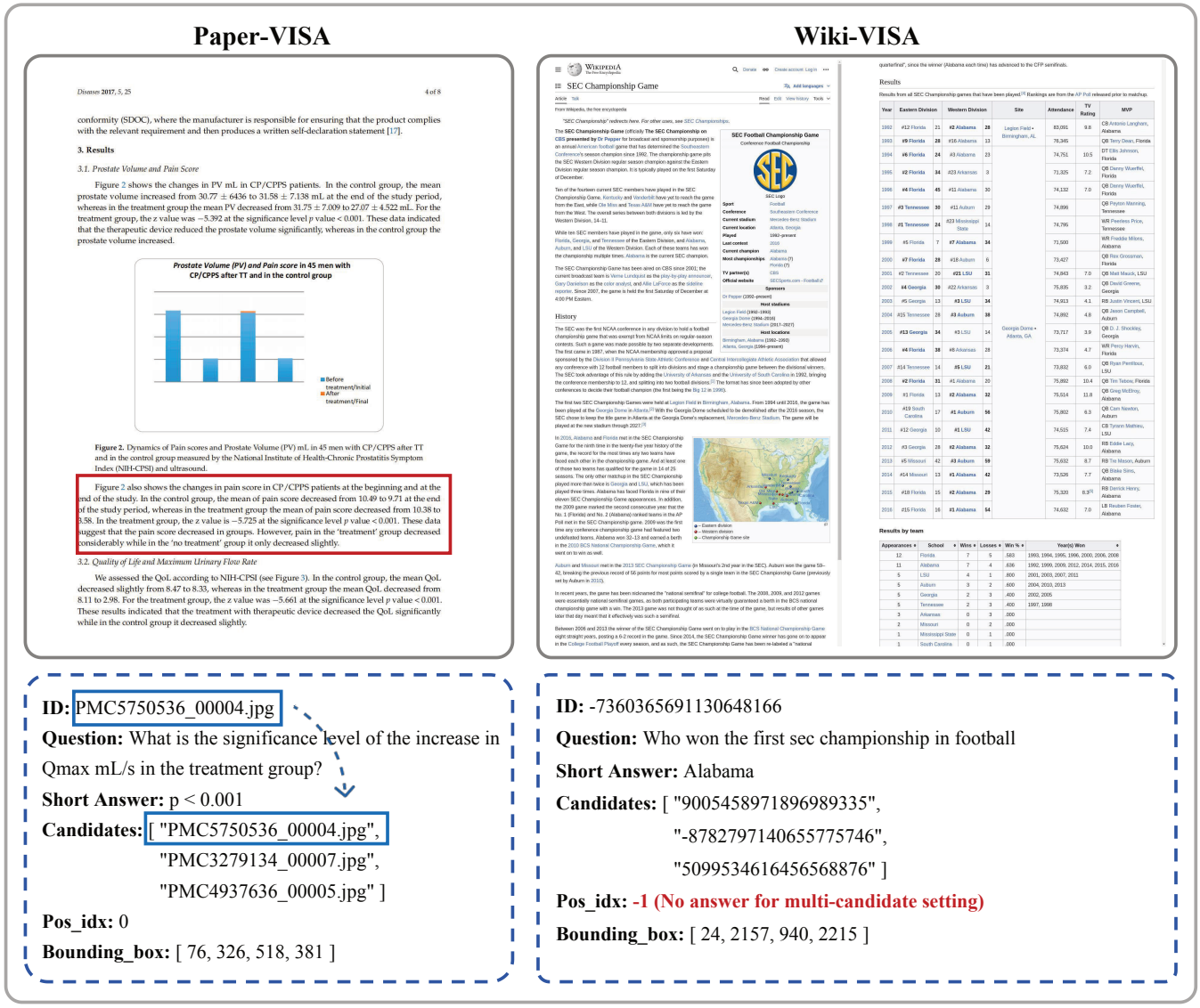
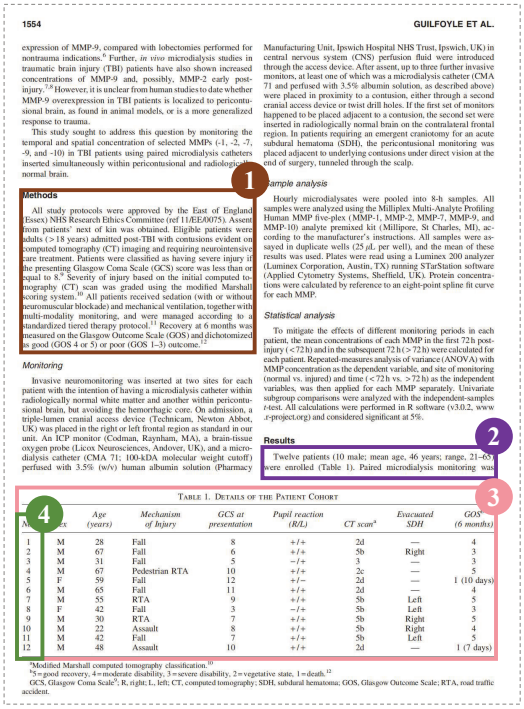


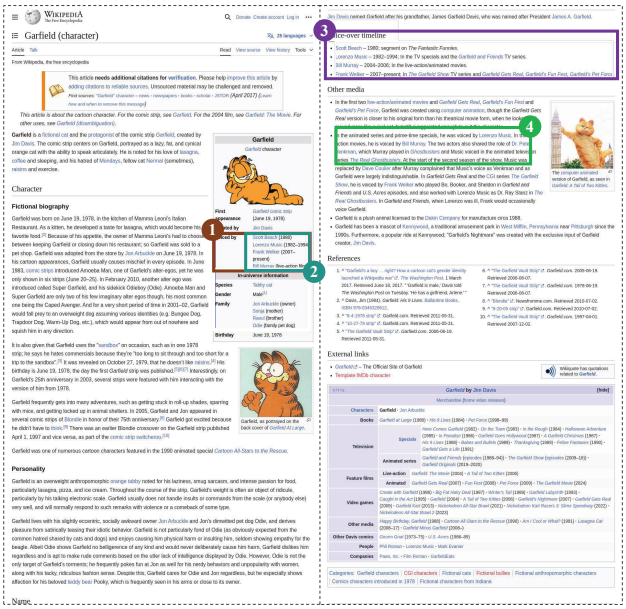
Figure 5: Data examples from Paper-VISA (left) and Wiki-VISA (right). Each image is assigned a unique identifier, with every dataset entry containing a reference image paired with a specific question, ground-truth answer, and answer source localized by a bounding box. In multi-image scenarios, a retriever selects two images and appends their IDs to the reference image, forming a candidate list. For example, the red bounding box (left) indicates the answer source, where `pos_idx=0` signifies that the reference image occupies the first position in the candidate list. For entries lacking ground-truth answers (right), the reference image is substituted with an irrelevant image in the candidate list (`pos_idx=-1`).



Question: How many patients were enrolled in the study?
Bounding Box: [62, 525, 552, 695]
Short Answer: 12

CoE Generation:
<think> 1. Scan the "Methods" section: This section describes how the study was conducted, including patient selection. However, it doesn't explicitly state the final number enrolled here.
{"bbox_2d": [60, 220, 298, 357], "Image_index": 1}
2. Scan the "Results" section: This section usually summarizes the findings, often starting with the number of participants. The first sentence of the "Results" section states: "Twelve patients (10 male; mean age, 46 years; range, 21–65) were enrolled (Table 1)."
{"bbox_2d": [312, 484, 552, 505], "Image_index": 1}
3. Examine Table 1: Table 1 is titled "DETAILS OF THE PATIENT COHORT"
{"bbox_2d": [62, 525, 552, 695], "Image_index": 1}
The table lists patients by number in the first column, **ranging from 1 to 12.**
{"bbox_2d": [62, 549, 85, 695], "Image_index": 1}
Both the text in the "Results" section and the entries in Table 1 confirm that 12 patients were enrolled. </think>
<answer> 12.
{"bbox_2d": [62, 525, 552, 695], "Image_index": 1} </answer>

Figure 6: The Chain-of-Evidence data generated from the Paper-VISA during the cold-start phase.



Question: Who is the voice of garfield the cat?
Bounding Box: [652, 674, 950, 781]
Short Answer: Lorenzo Music

CoE Generation:
<think> 1. Scan the document for sections related to the character's details.
2. The infobox on the right side typically contains such information.
I see a "Voiced by" field in the infobox.
{"bbox_2d": [652, 674, 943, 960], "Image_index": 1}
3. The infobox lists multiple names: Scott Beach, Lorenzo Music...
{"bbox_2d": [747, 674, 943, 781], "Image_index": 1}
4. There is a dedicated section titled "Voice-over timeline" which provides more detail. This section lists:
* Scott Beach; segment on *The Fantastic Funnies*.
* Lorenzo Music; In the TV specials and the *Garfield and Friends* TV series.
* Bill Murray; In the live-action/animated movies.
* Frank Welker; In *The Garfield Show* TV series and *Garfield Gets Real...
{"bbox_2d": [49, 1881, 950, 1998], "Image_index": 1}
5. The "Other media" section also mentions Lorenzo Music voiced Garfield in animated series and specials, while Bill Murray voiced him in the live-action movies.
{"bbox_2d": [49, 2153, 680, 2264], "Image_index": 1} </think>
<answer> According to the document, Garfield has been voiced by several actors: Scott Beach , Lorenzo Music, Frank Welker, and Bill Murray.
{"bbox_2d": [747, 674, 943, 781], "Image_index": 1},
{"bbox_2d": [49, 1881, 950, 1998], "Image_index": 1} </answer>

Figure 7: The Chain-of-Evidence data generated from the Wiki-VISA during the cold-start phase.

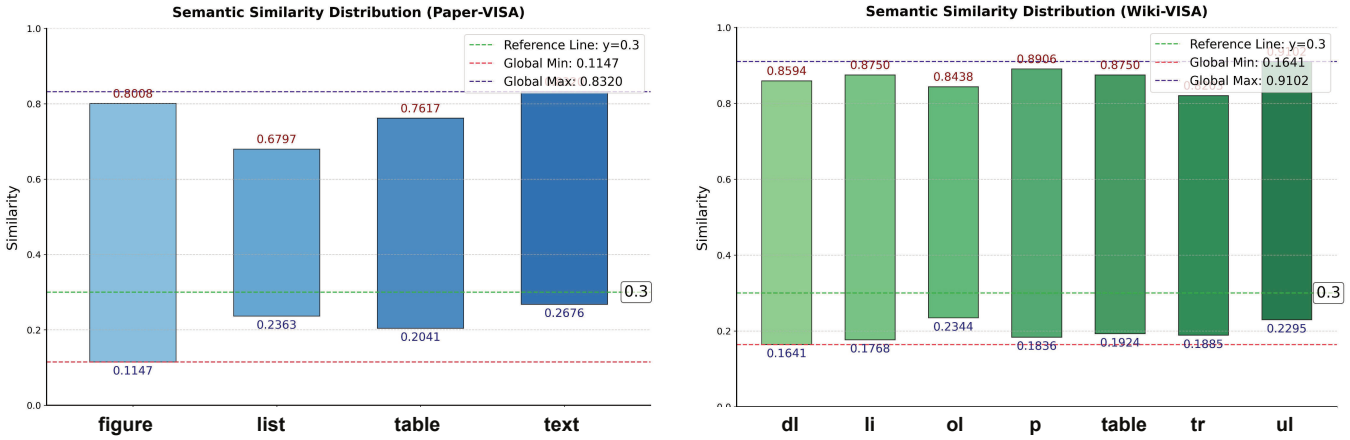


Figure 8: Semantic similarity distribution of synthetic CoE data. We computed the mean upper and lower bounds of semantic similarity between manually annotated evidence attributions and context-evidence pairs across different categories of synthetic CoE data in the cold-start stage. To ensure the effectiveness of the reward function, we selected a threshold $\tau = 0.3$ in Equation 8 to enforce well-aligned CoE reasoning steps during RL training.

<image> Image Size: ()
Task Description:
Given images of document pages, your task is to solve a document reasoning problem. Please first think about the long reasoning process, and then provide the user with the answer.
Restriction:
1. The reasoning process moves from the whole to the details, aligning with human observation.
2. While reasoning, the assistant needs to detect precise evidence for **each item** by using a 2D bounding box: For each item (such as figure, table, or factual information), the assistant should attach a 2D bounding box in "box_2d" and the image index number in "Image_index".
3. Finally, provide the final answer and locate the source of the answer via a 2D bounding box with the corresponding image index. If no answer is found in the images, please indicate "No answer." Here is an example:
Example: <example>
Question: <question>

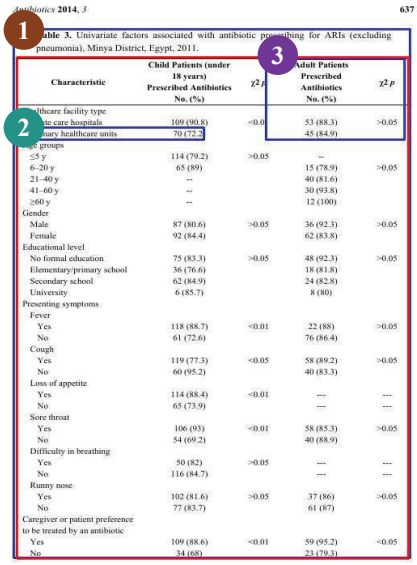
Figure 9: Prompt template used for CoE Generation with Gemini2.5 pro.

<|im_start|>user
<image> Image Size: ()
Task Description:
Given a document image and a relevant question, analyze the image to extract information relevant to the question, then provide the final answer. Finally, please locate the source of the final answer via a bounding box with an image index.
Restriction:
1. For each identified element (e.g., figures, tables, or factual text) during analysis, provide a bounding box and include its image index to highlight the visual evidence.
2. The analysis and final answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> analysis with visual evidence. </think><answer> the final answer and the corresponding bounding box as its source. </answer>
3. Each bounding box must be formatted as:
Bounding box: {bbox_2d: [x_1, y_1, x_2, y_2], image_index: i}
Question: <question> <|im_end|>
<|im_start|>assistant

Figure 10: Prompt template used for training and inference (Single-image) with Qwen2.5-VL. The structured prompt encourages attribution-aware reasoning.

<|im_start|>user
<image> Image Size: ()
<image> Image Size: ()
<image> Image Size: ()
Task Description:
Given document images and a relevant question, analyze the images to extract information relevant to the question, then provide the final answer. Finally, please locate the source of the final answer via a bounding box with an image index.
Restriction:
1. For each identified element (e.g., figures, tables, or factual text) during analysis, provide a bounding box and include its image index to highlight the visual evidence.
2. The analysis and final answer are enclosed within <think> </think> and <answer> </answer> tags, respectively, i.e., <think> analysis with visual evidence. </think><answer> the final answer and the corresponding bounding box as its source. </answer>
3. Each bounding box must be formatted as:
Bounding box: {bbox_2d: [x_1, y_1, x_2, y_2], image_index: i}
Question: <question> <|im_end|>
<|im_start|>assistant

Figure 11: Prompt template used for training and inference (Multi-image) with Qwen2.5-VL. The structured prompt encourages attribution-aware reasoning.

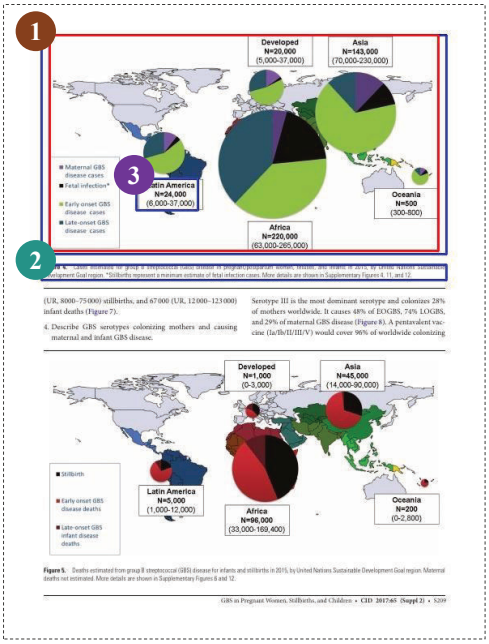


Question: How many adult patients were prescribed antibiotics at primary healthcare units?
Bounding Box: [55, 116, 539, 779]
Short Answer: 45

CoE Generation of LAT:

<think>
The question asks for the number of adult patients prescribed antibiotics at primary healthcare units. I need to find Table 3, which presents univariate factors associated with antibiotic prescribing, and look for the row corresponding to "Primary healthcare units" under the "Adult Patients Prescribed Antibiotics" column.
1. Locate Table 3. It is clearly labeled and positioned in the middle of the page.
{"bbox_2d": [50, 78, 540, 777], "Image_index": 1}
2. Find the row for "Primary healthcare units\". This is the second row under the "Characteristics" column.
{"bbox_2d": [58, 208, 285, 224], "Image_index": 1}
3. Look at the value in the "Adult Patients Prescribed Antibiotics" column for this row. The value is 45.
{\"bbox_2d\": [362, 117, 533, 224], "Image_index": 1} </think>
<answer> 45.
{"bbox_2d": [50, 78, 540, 777], "Image_index": 1}
</answer>

Figure 12: The Chain-of-Evidence reasoning process of LAT on the Paper-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 1)

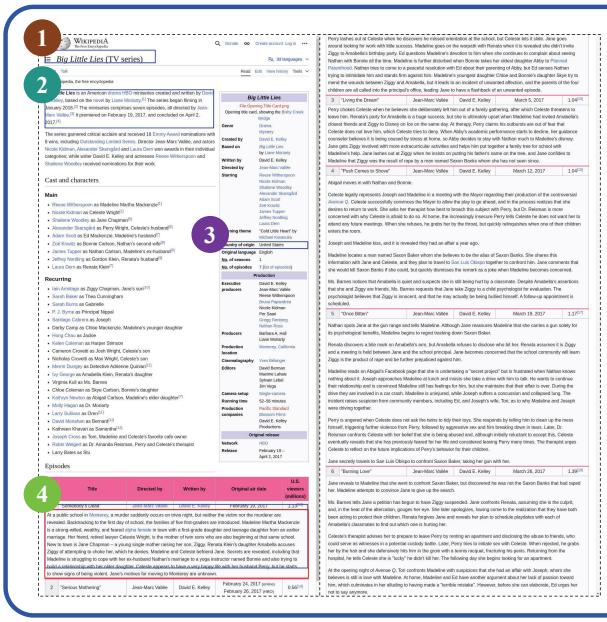


Question: What is the estimated number of GBS disease cases in Latin America in 2015?
Bounding Box: [57, 39, 537, 305]
Short Answer: 24,000

CoE Generation of LAT:

<think> The question asks about the estimated number of GBS disease cases in Latin America in 2015. I need to find the relevant figure or table that provides this information.
1. Scan the document for figures and tables related to GBS disease cases and Latin America. Figure 4 seems relevant.
{"bbox_2d": [46, 38, 546, 308], "Image_index": 1}
2. Examine Figure 4\'s caption and content. The caption mentions "Cases estimated for group B Streptococcal (GBS) disease in pregnant/postpartum women, fetuses, and infants in 2015".
{"bbox_2d": [46, 321, 546, 340], "Image_index": 1}
3. Look for specific data related to Latin America. The pie chart section labeled "Latin America" shows a value of 24,000 cases.
{"bbox_2d": [163, 215, 239, 254], "Image_index": 1}
Therefore, the estimated number of GBS disease cases in Latin America in 2015 is 24,000.
</think>
<answer> 24,000.
{"bbox_2d": [46, 38, 546, 308], "Image_index": 1}
</answer>

Figure 13: The Chain-of-Evidence reasoning process of LAT on the Paper-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 2)



Question: Where does the show big little lies take place?
Bounding Box: [24, 1661, 940, 1900]
Short Answer: Monterey, California

CoE Generation of LAT:

<think> The question asks about the location of the TV series "Big Little Lies". I need to find information related to the setting or filming locations within the provided image.

1. The title of the page is "Big Little Lies (TV series)".

{"bbox_2d": [24, 56, 407 100], "Image_index": 1}

2. The introductory paragraph mentions that the series began filming in January 2016. It also states that the series premiered on February 19, 2017, and concluded on April 2, 2017.

{"bbox_2d": [24, 194, 610, 327], "Image_index": 1}

3. The infobox on the right side of the page includes an entry for "Country of origin" which is listed as United States.

{"bbox_2d": [636, 724, 937, 748], "Image_index": 1}

4. The plot summary for the first episode mentions that the story takes place at a public school in Monterey.

{"bbox_2d": [24, 1640, 937, 1862], "Image_index": 1}

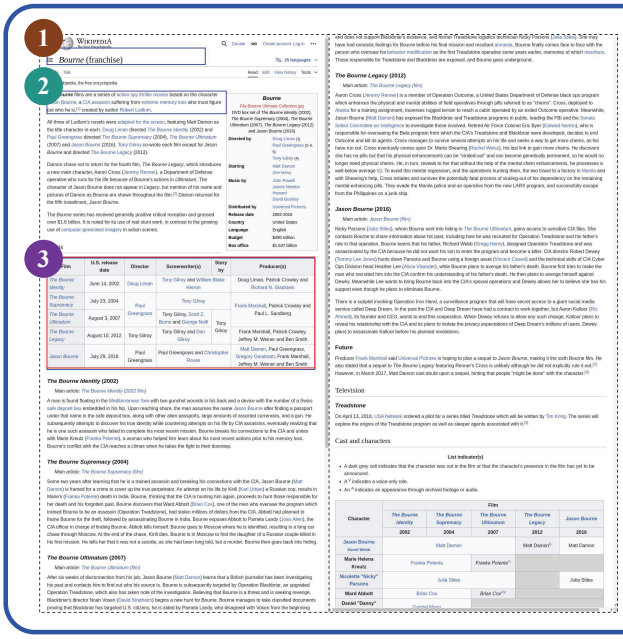
Based on these points, the show "Big Little Lies" takes place primarily in Monterey, California </think>

<answer>

Monterey, California.

{"bbox_2d": [24, 1640, 937, 1862], "Image_index": 1} </answer>

Figure 14: The Chain-of-Evidence reasoning process of LAT on the Wiki-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 3)



Question: When did the first jason bourne movie come out?
Bounding Box: [24, 758, 931, 1136]
Short Answer: June 14, 2002

CoE Generation of LAT:

<think> The question asks about the release date of the first Jason Bourne movie. I need to find this information in the provided document image.

1. The title of the page is "Bourne (franchise)".

{"bbox_2d": [24, 65, 279, 104], "Image_index": 1}

2. The introductory paragraph mentions that the films are based on the character Jason Bourne created by Robert Ludlum. It also states that all three of Ludlum's novels were adapted for the screen.

{"bbox_2d": [23, 194, 631, 400], "Image_index": 1}

3. The table lists the films chronologically. The first entry is \"The Bourne Identity\" with a U.S. release date of June 14, 2002.

{"bbox_2d": [24, 758, 932, 1135], "Image_index": 1}

Therefore, the first Jason Bourne movie came out on June 14, 2002. </think>

<answer>

June 14, 2002.

{"bbox_2d": [24, 758, 932, 1135], "Image_index": 1} </answer>

Figure 15: The Chain-of-Evidence reasoning process of LAT on the Wiki-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 4)

Question: How long can the isolation of corneal endothelial cell culture range from?
Bounding Box: -1
Short Answer: No Answer

CoE Generation of LAT:

<think> The question asks about the time range for isolating corneal endothelial cell cultures. I need to find the section in the provided images that discusses this process and identify the time frame mentioned.

1. Scan the images for keywords related to "corneal endothelial cells" and "isolation". Image 3 seems most relevant as it mentions culturing of endothelial cells on microcarriers.

2. Examine Image 3 more closely. The section titled "Culturing of endothelial cells on microcarriers" describes the process.

3. Look for information on the time range. The text states that the cells were cultured in DMEM media containing 10 % heat-inactivated fetal bovine serum (GIBCO, Gaithersburg, USA), 100 U/ml penicillin, 100 \u03bcg/ml streptomycin, and 2 mM glutamine. It also mentions that incubation was conducted at 37 \u00b0C in a 5 % CO2 humidified atmosphere. However, there is no explicit mention of how long the isolation process takes. The text focuses on the culture conditions after isolation, not the duration of the isolation itself. </think>

<answer> No answer.

</answer>

Figure 16: (Multi-image) The Chain-of-Evidence reasoning process of LAT on the Paper-VISA. This is a case in an unanswerable scenario. (Case 5)

Question: Where was the 3/5 compromise originally stated in the constitution ?
Bounding Box: [24, 820, 941, 872]
Short Answer: Article 1, Section 2, Clause 3

CoE Generation of LAT:

<think> The question asks about the original location of the Three-Fifths Compromise in the Constitution. I need to find the section of the document that discusses the Three-Fifths Compromise and identify where it is mentioned in the Constitution.

1. Scan the documents for the term \"Three-Fifths Compromise\". I can see this term in Document Image 3.
{"bbox_2d": [22, 64, 470, 101], "Image_index": 3}

2. Read the introduction of the Three-Fifths Compromise section in Document Image 3. It states that the Three-Fifths Compromise was reached during the 1787 United States Constitutional Convention.
{"bbox_2d": [22, 189, 941, 394], "Image_index": 3}

3. Locate the specific clause in the Constitution that mentions the Three-Fifths Compromise. The text says it's found in Article 1, Section 2, Clause 3.
{"bbox_2d": [22, 821, 941, 894], "Image_index": 3} </think>

<answer>
Article 1, Section 2, Clause 3 of the United States Constitution.
{"bbox_2d": [22, 821, 941, 894], "Image_index": 3} </answer>

Figure 17: (Multi-image) The Chain-of-Evidence reasoning process of LAT on the Wiki-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 6)