

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

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.

set to $5e-5$ 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} ($\text{EM} = 1$), 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 4xA800 (80 GB) GPUs,

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
RL Training	lora dropout	0.05
	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 time 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.

B Dataset Construction

VISA Datasets

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 Yipes 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.

Algorithm 1: LAT Framework

```

Input: Sampled query set  $\mathcal{Q}_{\text{cold\_start}}$  and  $\mathcal{Q}_{\text{RL}}$ 
Parameter: Learning rates  $\eta_1, \eta_2$ ; Reward thresholds  $\tau, \delta, \epsilon, \gamma$ 
Output: LAT model  $\pi_\theta$ 

1: Stage I: Cold-start Supervised Fine-tuning
2: for each query  $q \in \mathcal{Q}_{\text{cold\_start}}$  do
3:   Generate CoE reasoning traces using Gemini2.5 pro
4:   Filter outputs by the recall metric
5:   if Recall( $a, a_{gt}$ )  $\geq \gamma$  then
6:      $\mathcal{D}_{\text{final}} \leftarrow \mathcal{D}_{\text{final}} \cup \{q\}$ 
7:   end if
8: end for
9: Manually verify bounding boxes and format
10: Train model  $\pi_\theta$  via SFT on verified data  $\mathcal{D}_{\text{final}}$  with learning rate  $\eta_1$ 
11: Stage II: Reinforcement Learning with GRPO
12: for each training step do
13:   Sample a batch of queries  $q \sim \mathcal{Q}_{\text{RL}}$ 
14:   Model  $\pi_\theta$  generates CoE reasoning steps  $\{r_1, \dots, r_T\}$  and bounding boxes  $\{B_1, \dots, B_T\}$  for each query  $q$ 
15:   Compute reward  $R = R_{\text{acc}} + R_{\text{step}} + R_{\text{ground}} + R_{\text{format}}$ 
16:   Update policy  $\pi_\theta$  using GRPO with  $R$  and learning rate  $\eta_2$ 
17: end for
18: return final LAT model  $\pi_\theta$ 

```

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
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.

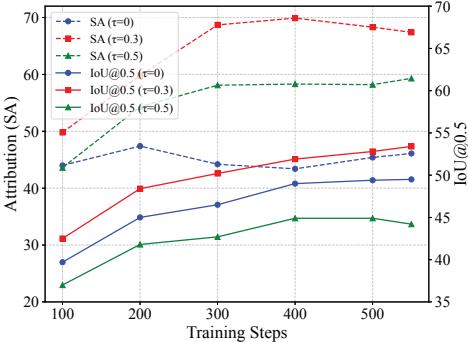


Figure 4: Performance variation with different τ (Wiki).

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 HuggingFace¹. 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 Cross-Task Adaptability

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 _{↑6.7}	10.4 _{↓1.1}
Wiki→Paper (Multi)	SFT	34.3	5.5
	LAT-Ind.	44.3 _{↑10.0}	21.9 _{↑16.4}

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

Dateset	Method	EM	IoU@0.5	SA
Paper	LAT-Full (Single)	46.2	48.4	33.8
	LAT-Full (Multi)	46.0	47.1	33.6
Wiki	LAT-Full (Single)	73.1	57.8	59.6
	LAT-Full (Multi)	72.8	52.3	51.3

Table 10: LAT retains single-image performance after multi-image training, demonstrating cross-task generalization.

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 9). These results indicate that LAT improves transfer effectiveness while retaining adaptability across diverse document types. Table 10 further confirms that multi-image training preserves single-image performance, demonstrating exceptional cross-task generalization without interference.

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 11 reports the accuracy of correctly detecting no-answer cases in multi-image settings.

G Limitations

While LAT substantially improves stepwise attribution and overall performance, several practical aspects remain open for future enhancement.

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

Reward Thresholds and Adaptivity. The threshold settings in the attribution recall reward are currently informed by annotated examples. To further stabilize training, future work will explore adaptive or soft reward formulations, where values exceeding the threshold contribute proportionally to the overall reward, thereby encouraging stronger evidence alignment.

Toward Multi-hop CoE Reasoning in VD-RAG. Existing multi-hop VD-RAG benchmarks primarily provide page-level evidence and are mainly designed for evaluation rather than supervision. Therefore, there is still a lack of datasets that can effectively support multi-hop reasoning, especially those enabling fine-grained evidence supervision.

Fine-grained Visual Attribution. The current visual attribution relies on bounding-box detection, which may overlook subtle semantic cues within charts or tables. Incorporating segmentation-level grounding can achieve higher spatial fidelity and richer evidence alignment.

Paper-VISA

Diseases 2017, 5, 25

4 of 8

conformity (SDOC), where the manufacturer is responsible for ensuring that the product complies with the relevant requirement and then produces a written self-declaration statement [17].

3. Results

3.1. Prostate Volume and Pain Score

Figure 2 shows the changes in PV mL in CP/CPPS patients. In the control group, the mean prostate volume increased from 30.7 ± 6.636 to 31.58 ± 7.138 mL at the end of the study period, whereas in the treatment group the mean PV decreased from 31.75 ± 7.009 to 27.07 ± 4.522 mL. For the treatment group, the z value was -5.392 at the significance level $p < 0.001$. These data indicated that the therapeutic device reduced the prostate volume significantly, whereas in the control group the prostate volume increased.

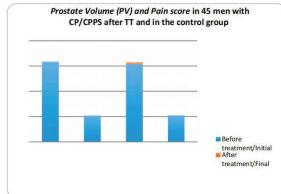


Figure 2. Dynamics of Pain scores and Prostate Volume (PV) mL in 45 men with CP/CPPS after TT and in the control group measured by the National Institute of Health-Chronic Prostatitis Symptom Index (NIH-CPSI) and ultrasound.

Figure 2 also shows the changes in pain scores in CP/CPPS patients at the beginning and at the end of the study. In the control group, the mean of pain score decreased from 10.38 ± 3.58 to 8.38 at the end of the study period, whereas in the treatment group the mean of pain score decreased from 10.38 ± 3.58 to 8.38 . In the treatment group, the z value was -5.725 at the significance level $p < 0.001$. These data suggest that the pain score decreased in groups. However, pain in the 'treatment' group decreased considerably while in the 'no treatment' group it only decreased slightly.

3.2. Quality of Life and Maximum Urinary Flow Rate

We assessed the QoL according to NIH-CPSI (see Figure 3). In the control group, the mean QoL decreased slightly from 8.47 ± 8.33 , whereas in the treatment group the mean QoL decreased from 8.11 ± 2.98 . In the treatment group, the z value was -5.661 at the significance level $p < 0.001$. These data indicated that the treatment with therapeutic device decreased the QoL significantly while in the control group it decreased slightly.

Wiki-VISA

Q: Diseases > SEC Championship Game
A: SEC Championship Game

The SEC Championship Game (often The SEC Championship on CBS presented by Dr Pepper for broadcast and sponsorship purposes) is an annual college football game between the two highest ranked Conference's season champion since 1996. The championship game pits the SEC West Division regular season champion against the Eastern Division regular season champion. It is typically played on the first Saturday of December.

Ten of the former current SEC members have played in the SEC Championship Game, while the other five have not. The first game from the East, while the Miss and Texas A&M have yet to reach the West. The Jordan series between both divisions is held at the Georgia Dome in Atlanta.

While ten SEC members have played in the game, only four have won: Florida, Georgia, and Tennessee of the Eastern Division, and Alabama, Auburn, and Mississippi State of the Western Division.

The conference multiple times. Alabama is the current SEC champion.

The SEC Championship Game has been aired on CBS since 2001; the current broadcast home is Verne Lundquist at the play-by-play announcer, Greg Gosselin at the color analyst, and Lee Corso at the pre-game reporter. Since 2007, the game is held on the first Saturday of December at 4:00 PM Eastern.

History

The SEC began in 1912. Conference's first decision was to hold a football championship game that was exempt from NCAA rules on football contests. Such a game was made possible by two key research developments:

The first came in 1987, when the NCAA membership approved a proposal sponsored by the University of Michigan and the University of Wisconsin-Madison to allow intercollegiate athletic association that allowed any conference with 12 football members to split into divisions and stage a championship game between the division winners.

The SEC took advantage of this rule by adding the University of Arkansas and the University of South Carolina in 1992, bringing the conference to 12 members.

The SEC then decided to change its name to the Southeastern Conference to reflect their football conference (the Big 12 is 2009).

The first two SEC Championship Games were held at Legion Field in Birmingham, Alabama. From 2004 until 2010, the game has been held at the Georgia Dome in Atlanta. The Georgia Dome stadium scheduled to be demolished after the 2010 season, the SEC chose to keep the title game in Atlanta at the Georgia Dome stadium, extension first. Starting the game will be played at the new stadium (Georgia 2012).

In 2011, the SEC moved the SEC Championship Game to the Georgia Dome in Atlanta for the second consecutive year.

For the next 10 years, the SEC has been the only conference to have a football championship game that is held in the same city every year.

The record for the most times any two teams have faced each other in the SEC Championship Game is 12, which is one of those two teams that qualified for the game in 2015.

The only other institution in the SEC Championship played more than once in the SEC Championship Game is the University of Alabama, which has faced the University of Tennessee twice in the game.

In 2009, the SEC announced that the 2010 game would be the last time any conference championship game had featured two undefeated teams. Alabama won 32–10 and earned both in the 2010 SEC Championship Game, which went on to win the game.

Alabama and Missouri met in the 2013 SEC Championship Game (in Missouri's 2nd year in the SEC). Alabama won the game $59-34$.

Missouri had the most points scored by a single team in the SEC Championship Game in 2013.

In recent years, the game has been renamed the "National semifinal" for college football.

The 2008, 2009, and 2010 games were virtually guaranteed a berth in the BCS national championship game, while the 2011 game was not enough to earn a spot in the game, but most of other years later that day meant that it effectively won such a semifinal.

Between 2006 and 2013 the winner of the SEC Championship Game went on to play in the BCS National Championship Game eight straight years, winning 6 times in the game. Since 2014 the SEC Championship Game has been moved to appear in the College Football Playoff every season, and as such the SEC Championship Game has never been named a "semifinal".

quarterfinal", since the winner (Alabama each time) has advanced to the CFP semifinals.

Results from all SEC Championship games that have been played^[1] Rankings are from the AP Poll released prior to match.

Year Eastern Division Western Division Site Attendance TV MHP

1990 #22 Florida 21 #2 Alabama 28 Legion Field* Birmingham, AL 83,001 9.8

1991 #9 Florida 28 #8 Alabama 13 DT Fins, Jacksonville, Florida 76,345 10.5

1992 #6 Florida 24 #3 Alabama 23 74,751 10.5

1993 #2 Florida 34 #23 Arkansas 9 QB Dan Wuerffel, Florida 71,305 7.2

1994 #4 Florida 45 #3 Alabama 30 74,332 7.0

1995 #9 Tennessee 39 #12 Auburn 29 74,006 7.0

1996 #1 Tennessee 24 #23 Mississippi State 34 74,795 7.0

1997 #7 Florida 7 #7 Alabama 34 73,500 7.0

1998 #9 Florida 28 #18 Auburn 6 73,407 7.0

1999 #7 Tennessee 20 #21 LSU 31 74,943 7.0

2000 #4 Georgia 30 #22 Arkansas 8 75,805 3.2

2001 #5 Georgia 34 #1 LSU 34 74,913 4.2

2002 #5 Georgia 28 #12 Auburn 38 74,892 4.6

2003 #13 Georgia 34 #3 LSU 34 Georgia Dome* Atlanta, GA 73,717 3.9

2004 #4 Georgia 38 #8 Arkansas 28 73,374 4.7

2005 #13 Georgia 30 #8 Auburn 28 73,855 7.4

2006 #4 Georgia 10 #1 LSU 42 74,001 6.0

2007 #13 Tennessee 14 #5 LSU 23 73,832 6.0

2008 #2 Florida 33 #5 Alabama 20 75,802 10.4

2009 #3 Florida 13 #7 Alabama 32 75,514 11.8

2010 #13 Florida 17 #2 Auburn 56 75,802 6.3

2011 #12 Georgia 10 #1 LSU 42 74,555 7.4

2012 #3 Georgia 28 #8 Alabama 32 75,624 10.8

2013 #5 Missouri 42 #3 Auburn 59 75,652 8.7

2014 #5 Missouri 13 #4 Alabama 42 73,206 7.7

2015 #8 Florida 15 #3 Alabama 23 75,320 8.9

2016 #5 Florida 18 #3 Alabama 54 74,602 7.6

LB Reuben Foster, Alabama

Results by team

Appearances • School • Wins • Losses • Win % • Year(s) Won

12 Florida 7 5 58% 1993, 1994, 1995, 1996, 2000, 2006, 2008

12 Alabama 7 4 63% 1992, 1993, 2006, 2012, 2014, 2015, 2016

5 LSU 4 2 50% 2001, 2002, 2007, 2012

5 Auburn 3 2 66% 2004, 2010, 2013

5 Georgia 2 3 40% 2002, 2005

5 Tennessee 2 3 40% 1997, 1998

3 Arkansas 0 3 00% 2003, 2004, 2005

2 Missouri 0 2 00% 2000, 2001

1 Mississippi State 0 1 00% 2003

1 South Carolina 0 1 00%

ID: [PMC5750536_00004.jpg]

Question: What is the significance level of the increase in Qmax mL/s in the treatment group?

Short Answer: $p < 0.001$

Candidates: ["PMC5750536_00004.jpg",
"PMC3279134_00007.jpg",
"PMC4937636_00005.jpg"]

Pos_idx: 0

Bounding_box: [76, 326, 518, 381]

ID: -7360365691130648166

Question: Who won the first sec championship in football

Short Answer: Alabama

Candidates: ["9005458971896989335",
"-8782797140655775746",
"5099534616456568876"]

Pos_idx: -1 (No answer for multi-candidate setting)

Bounding_box: [24, 2157, 940, 2215]

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 blue 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).

Age	Sex	Age (years)	Mechanism of Injury	GCS at presentation	Pupil reaction (R/L)	CT scan*	Evacuated	SDH	GOS (6 months)
4					4	5	6	7	8
1	M	28	Fall	8	+/-	2d	—	—	4
2	M	67	Fall	6	+/-	5b	Right	—	5
3	M	31	Fall	7	+/-	3	—	—	3
4	M	47	Pedestrian RTA	10	+/-	5b	Left	—	1 (10 days)
5	F	59	Fall	12	+/-	2d	—	—	5
6	M	65	Fall	11	+/-	2d	—	—	5
7	M	54	RTA	9	+/-	5b	Left	—	5
8	F	42	Fall	3	+/-	5b	Left	—	3
9	M	30	RTA	7	+/-	5b	Right	—	5
10	M	23	Assault	8	+/-	5b	Right	—	4
11	M	42	Fall	7	+/-	5b	Left	—	5
12	M	48	Assault	10	+/-	2d	—	—	1 (7 days)

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

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

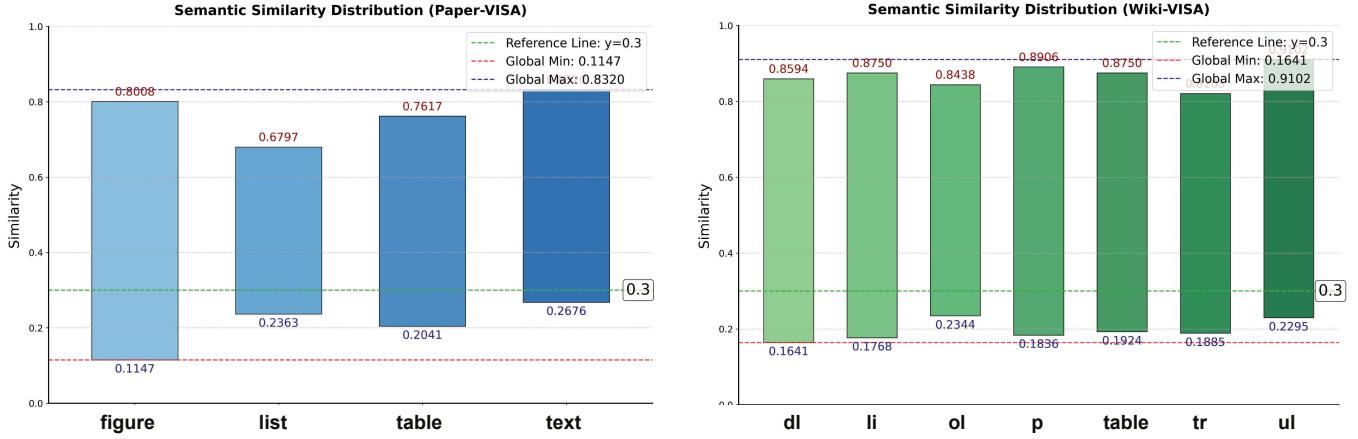


Figure 8: Semantic similarity distribution of synthetic CoE data. We computed the mean upper and lower bounds of semantic similarity between manually annotated evidence attributions and context-evidence pairs across different categories of synthetic CoE data in the cold-start stage. To ensure the effectiveness 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: [x₁, y₁, x₂, y₂], 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: [x₁, y₁, x₂, y₂], 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.

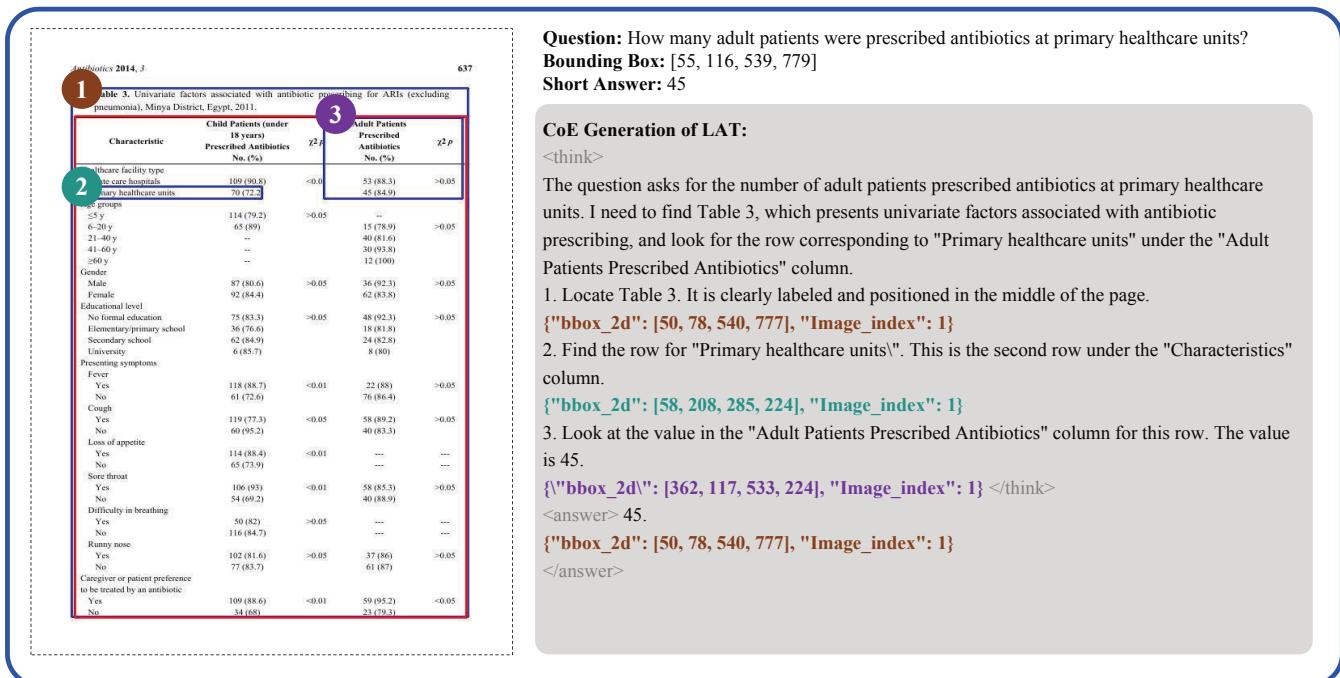


Figure 12: The Chain-of-Evidence reasoning process of LAT on the Paper-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 1)

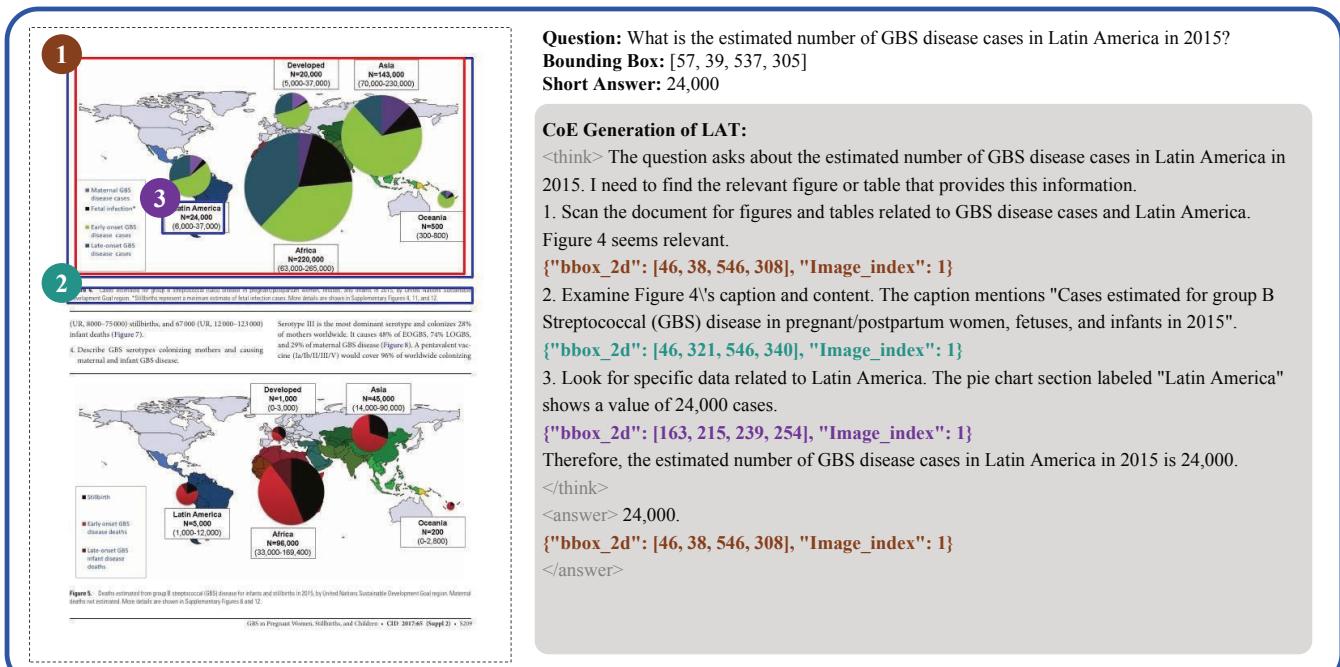


Figure 13: The Chain-of-Evidence reasoning process of LAT on the Paper-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 2)

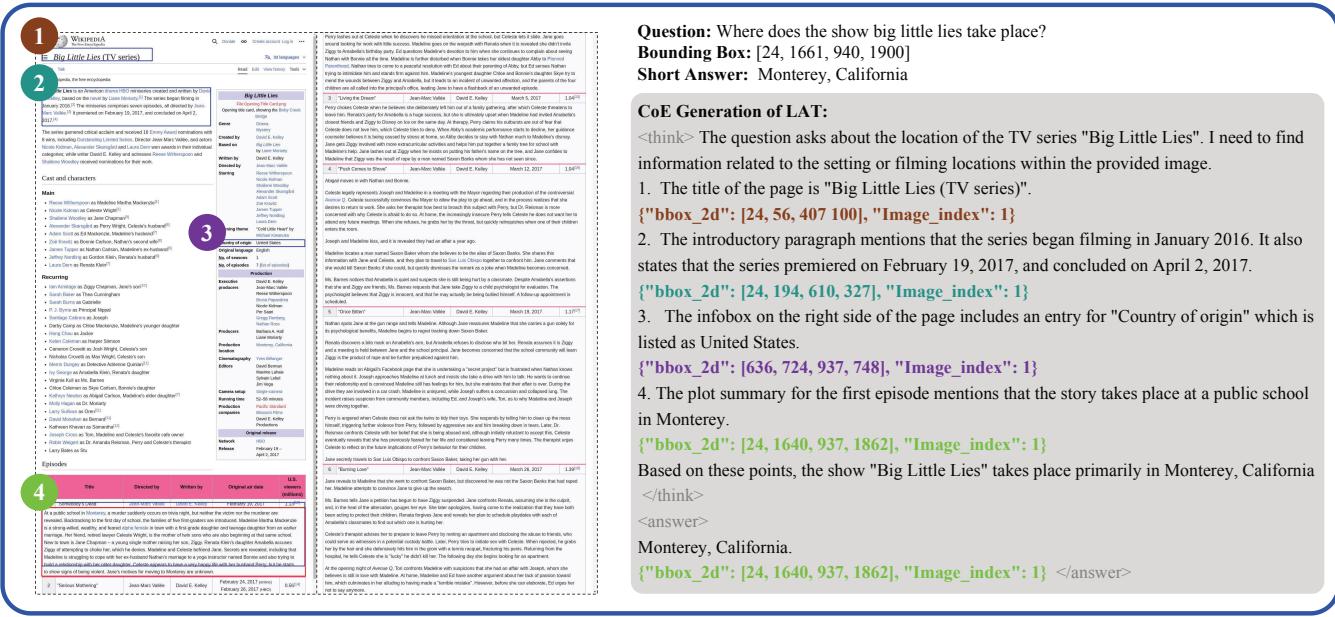


Figure 14: The Chain-of-Evidence reasoning process of LAT on the Wiki-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 3)

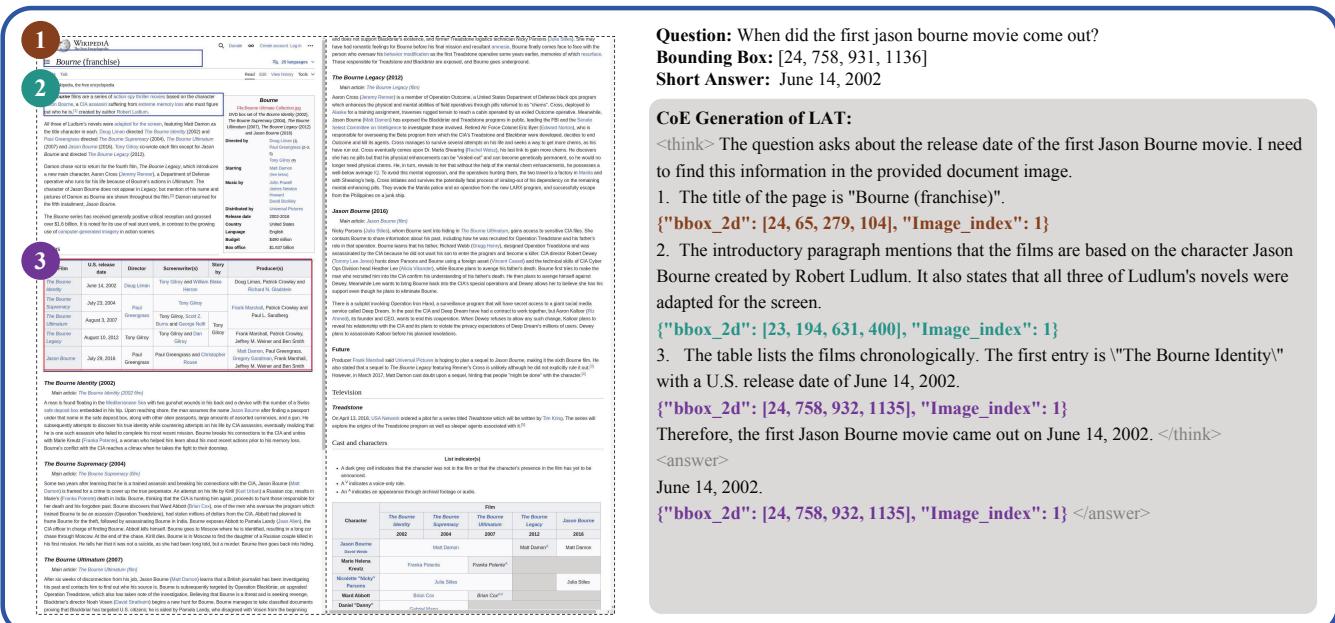


Figure 15: The Chain-of-Evidence reasoning process of LAT on the Wiki-VISA. Blue boxes represent model predictions, while the red box indicates the ground truth bounding box from the dataset. (Case 4)

