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

Shuochen Liu, Pengfei Luo, Chao Zhang, Yuhao Chen, Haotian Zhang, Qi Liu, Xin Kou, Tong Xu*, Enhong Chen

University of Science and Technology of China
shuochenliu@mail.ustc.edu.cn, tongxu@ustc.edu.cn

Abstract

Aiming to identify precise evidence sources from visual documents, visual evidence attribution for visual document retrieval–augmented generation (VD-RAG) ensures reliable and verifiable predictions from vision-language models (VLMs) in multimodal question answering. Most existing methods adopt end-to-end training to facilitate intuitive answer verification. However, they lack fine-grained supervision and progressive traceability throughout the reasoning process. In this paper, we introduce the **Chain-of-Evidence (CoE)** paradigm for VD-RAG. CoE unifies Chain-of-Thought (CoT) reasoning and visual evidence attribution by grounding reference elements in reasoning steps to specific regions with bounding boxes and page indexes. To enable VLMs to generate such evidence-grounded reasoning, we propose **Look As You Think (LAT)**, a reinforcement learning framework that trains models to produce verifiable reasoning paths with consistent attribution. During training, LAT evaluates the attribution consistency of each evidence region and provides rewards only when the CoE trajectory yields correct answers, encouraging process-level self-verification. Experiments on vanilla Qwen2.5-VL-7B-Instruct with Paper- and Wiki-VISA benchmarks show that LAT consistently improves the vanilla model in both single- and multi-image settings, yielding average gains of 8.23% in soft exact match (EM) and 47.0% in IoU@0.5. Meanwhile, LAT not only outperforms the supervised fine-tuning baseline, which is trained to directly produce answers with attribution, but also exhibits stronger generalization across domains.

Code — https://github.com/PolarisLiu1/LAT
Extended version — https://arxiv.org/pdf/2511.12003

1 Introduction

With the development of multimodal understanding capabilities in vision-language models (VLMs) (Chen et al. 2025c; Bai et al. 2025a), visual document retrieval-augmented generation (VD-RAG) has emerged as a critical research frontier. Nevertheless, current VLMs remain susceptible to hallucinations, whereby their outputs may deviate from the source document content (Bai et al. 2025b). Without reliable visual evidence attribution mechanisms to identify sources in documents, users cannot intuitively trace back the specific information employed by the model, thereby reducing the reliability of VD-RAG systems in applications.

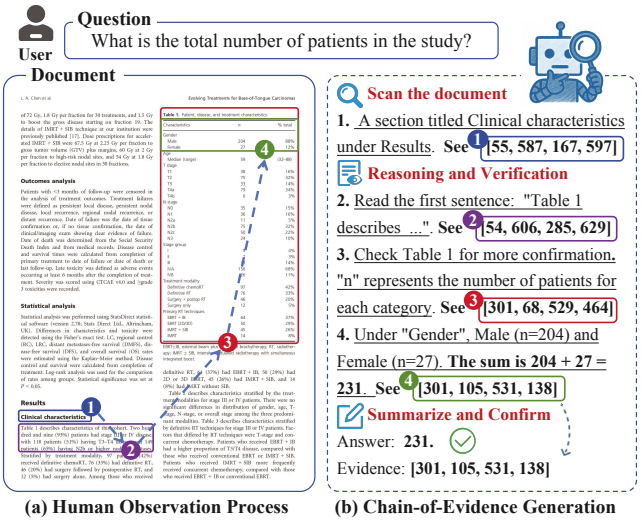


Figure 1: (a) Humans infer information by observing and locating supporting evidence in the document. (b) Each element in the reasoning step is linked to a visual attribution via a bounding box during Chain-of-Evidence generation.

Along this line, recent works (Shao et al. 2024a; Wu et al. 2025; Qi et al. 2025) mitigate hallucination in single-image Chain-of-Thought (CoT) (Kojima et al. 2022) reasoning by attending to critical visual regions as evidence, leveraging large-scale annotations of stepwise evidence regions. VISA (Ma et al. 2024b) further incorporates visual evidence attribution into the VD-RAG framework by linking answers with supporting sources via bounding boxes. Despite these advances, effective evidence attribution in visual documents still faces challenges: **(1) Lack of progressive reasoning mechanisms for verifiable attribution.** Existing methods, such as VISA, directly associate answers with final evidence, but do not reveal the intermediate process that would help users clearly understand the content and trace how the evidence is located. In contrast, when addressing complex problems, humans do not directly locate evidence but progressively search for relevant information. As shown

*Corresponding author.

in Figure 1a, such an observation pathway is structured as "chapter-paragraph summary-specific element", localizing evidence from coarse to fine. However, current VLMs struggle to replicate this progressive observation process, especially in multi-image scenarios. **(2) Limited supervision for learning multimodal reasoning.** Training models to learn CoT reasoning requires extensive annotated data (Xu et al. 2025; Wang et al. 2025), especially for stepwise evidence attribution. However, manually constructing such datasets is a costly process. Therefore, how to effectively learn stepwise visual evidence attribution and generalize reasoning abilities under limited annotated data remains a critical challenge.

Given these challenges, we introduce **Chain-of-Evidence (CoE)**, a reasoning paradigm that integrates Chain-of-Thought (CoT) reasoning with visual evidence attribution, designed for VD-RAG. Unlike existing methods that perform reasoning within a single image or directly link the final answer to the source, CoE models the intermediate reasoning trajectory by grounding each step to a supporting source from documents. This coarse-to-fine attribution process mirrors human problem-solving and enhances the reliability of VD-RAG. To this end, we introduce **Look-As-You-Think (LAT)**, a two-stage training approach designed to implement the proposed CoE paradigm. Specifically, in the first stage, we fine-tune VLM on a set of few human-verified annotated CoE data to learn reasoning patterns. Then we adopt reinforcement learning (RL) under the Group Relative Policy Optimization (GRPO) (Shao et al. 2024b) through a tailored reward design. The model is guided by a stepwise reward based on the semantic alignment between predicted visual evidence and corresponding context, encouraging faithful reasoning without requiring stepwise annotations. Furthermore, the reward is issued only when the CoE trajectory yields the correct answer. We further incorporate the outcome reward to ensure answer accuracy and guide the model in defining the scope of answer localization. This combined reward scheme enhances the ability to generate CoE reasoning, thereby linking sub-step verification to end-task performance. Our contributions are summarized as follows:

1) In the VD-RAG scenario, we formalize the **Chain-of-Evidence (CoE)** paradigm by modeling multimodal reasoning as a sequence of grounded steps, where reference elements (e.g., figures, tables, or factual information) are linked to their source through a bounding box and page index.

2) Building upon CoE, we propose **LAT**, an RL-based approach that jointly optimizes reasoning and visual grounding through a stepwise reward. By aligning each reference element with its visual evidence, LAT enables attribution-aware reasoning with few CoE-annotated samples.

3) LAT balances traceability and performance. Compared to the vanilla model, it shows improvements of 8.23% EM and 47.0% IoU@0.5 in both single- and multi-image scenarios. LAT exhibits stronger cross-domain generalization.

2 Related Work

Visual Evidence Attribution

Early end-to-end grounding methods integrate object detection into generated text by using markdown hyperlinks to generate bounding-box tokens (Chen et al. 2023; Peng et al. 2023). They are trained on large corpora of grounded images and texts. Building on this foundation, multi-step visually grounded CoT frameworks (Shao et al. 2024a; Li et al. 2025; Wu et al. 2025; Xia et al. 2025) interleave reasoning and localization by predicting bounding boxes as evidence within the reasoning process, thereby yielding interpretable reasoning traces. However, these approaches have been validated exclusively on general visual perception tasks and rely on large-scale annotated evidence regions. They also do not explore the visual evidence attribution task in VD-RAG involving heterogeneous layouts and multi-page retrieval.

Moreover, existing text-based evidence attribution methods (Gao et al. 2023; Ye et al. 2024) in document RAG often operate at the document level, requiring users to read entire documents to locate supportive content. VISA (Ma et al. 2024b) first adapts visual evidence attribution to document screenshots by aligning the final answer with its evidence. Despite enabling intuitive verification of correctness, it fails to explicate the intermediate reasoning steps through which the model arrives at the answer. These limitations motivate the CoE reasoning paradigm, which generates faithful reasoning steps and validates each reference element against its corresponding source, as shown in Figure 1b.

RL for VLM Reasoning

Recent research has shown that RL-based policy optimization (Zhang et al. 2024) can improve the reasoning capabilities of large language models (LLMs) (Chen et al. 2025a,b; Zhu et al. 2024). DeepSeek-R1 (Guo et al. 2025) demonstrated that RL training elicits emergent CoT behaviors in LLMs, revealing the hallmark "aha moment". Inspired by this phenomenon, several methods (Peng et al. 2025) have extended R1-style RL strategies to VLMs, leveraging rule-based reward functions to boost performance on mathematical reasoning and visual perception tasks. Unlike supervised fine-tuning (SFT), RL-based approaches achieve deeper reasoning and stronger generalization without relying on extensive human-annotated data (Chu et al. 2025).

However, existing RL frameworks for multimodal tasks are primarily optimized for answer accuracy as the reward signal (Yang et al. 2025; Shen et al. 2025), with no explicit supervision for verifying intermediate reasoning (Ni et al. 2025; Cao et al. 2025), and without design considerations for the visual evidence attribution task in VD-RAG. Drawing on human reading strategies, we introduce a stepwise, process-level reward that aligns each reasoning step with verifiable evidence. Leveraging the extracted CoE, we explicitly reward trajectories that are evidentially consistent and culminate in the correct answer, thereby ensuring valid reasoning and fostering faithful, attribution-based explanations.

3 Proposed Approach

To remedy the lack of verifiable progressive reasoning for visual evidence attribution, we first formalize the **Chain-of-Evidence (CoE)** paradigm. Building upon CoE, we present **LAT**, a two-stage RL-based framework shown in Figure 2. Stage I performs supervised fine-tuning to align annotated

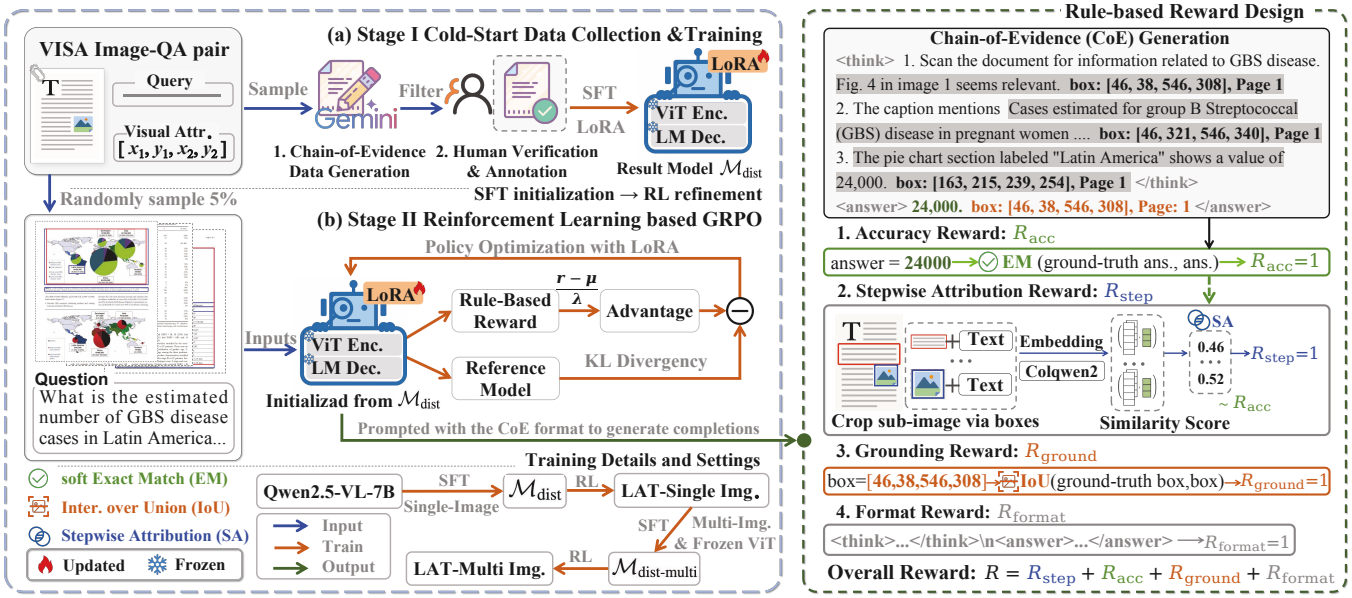


Figure 2: Overview of the proposed LAT framework. Left: A two-stage training pipeline. Stage I generates and filters the CoE data for fine-tuning. Stage II: The model undergoes refinement via RL under the GRPO algorithm. Right: Rule-based reward design. In GRPO training, the model generates CoE reasoning to guide policy updates through the reward signals.

CoE traces, and Stage II conducts fine-grained reinforcement learning with an answer-conditioned attribution reward that refines step-level grounding.

CoE Formalization and Notations

We formalize the generation with Chain-of-Evidence (CoE) reasoning as follows. Specifically, we define a textual query q and a set of document pages $\mathcal{P} = \{p_n\}_{n=1}^{N}$, which are pre-retrieved from the corpus. Given $(q, \mathcal{P})$, CoE requires a VLM ϕ to perform CoT reasoning with stepwise visual attribution and then produce both an answer and its supporting evidence for evaluation, formulated as:

$$\mathcal{R}, \mathcal{B}, \mathcal{A} = \phi(q, \mathcal{P}). \tag{1}$$

Here, $\mathcal{R} = \{r_t\}_{t=1}^{T}$ denotes the textual reasoning steps, while $\mathcal{B} = \{(i_t, B_t)\}_{t=1}^{T}$ is the corresponding evidence chain, where $i_t \in [1, N]$ indicates the page index and $B_t = [(x_1^t, y_1^t), (x_2^t, y_2^t)]$ specifies the bounding box of the visual evidence for r_t on i_t-th page. After CoE reasoning, the final output $\mathcal{A} = \{a, (i^*, B_{ans})\}$ consists of the answer a, the most relevant page $p_{i^*} \in \mathcal{P}$, and the bounding box $B_{ans} \in \mathcal{B}$ confirming the evidence that supports a. It should be noted that not all reasoning steps in the response require a bounding box over images, such as calculation or conclusion steps. We omit these cases in our formulation for mathematical conciseness.

Stage I: Cold-Start Data Collection and Training

To prime the model for CoE reasoning, we first sample $1,000$ instances from each training dataset and prompt a stronger proprietary model Gemini 2.5 Pro (Comanici et al. 2025) using two CoE exemplars as in-context demonstrations. The model produces stepwise rationales, each annotated with bounding-box evidence, yielding a cold-start corpus that reflects the desired reasoning format.

For each query q and its corresponding response y, we assess answer quality with a recall metric,

$$\text{Recall}(a, a_{gt}) = \frac{|a \cap a_{gt}|}{|a_{gt}|}, \tag{2}$$

where a_{gt} is the ground-truth answer from the dataset, $|a|$ denotes the number of words in the extracted answer portion a from the response y, and $|a \cap a_{gt}|$ is the number of overlapping words between a and a_{gt}. Only samples with recall above a threshold γ are retained to ensure sufficient answer accuracy in the initial CoE traces.

We manually verify and correct any bounding-box drift, retaining verified samples with correct answers ($\sim$30%) to ensure data quality. Details of the resulting dataset $\mathcal{D}_{\text{final}}$ used in the cold-start training, including its split distribution, are provided in Table 6 in the appendix. Next, we fine-tune the VLM on $\mathcal{D}_{\text{final}}$ using LoRA (Hu et al. 2022), aiming to minimize the cross-entropy loss between the generated output and the annotated reasoning sequences through SFT. The resulting model is defined as $\mathcal{M}_{\text{dist}}$.

Stage II: Unified Reasoning and Visual Attribution via Reinforcement Learning

To emulate the human observation process shown in Figure 1, the model needs to deliver both an accurate answer and attribution-aware CoE reasoning to identify evidence supporting the final answer. We decompose the overall objective into four sub-goals: answer accuracy, stepwise visual attribution quality, evidence grounding precision, and adherence to the structured output format. Accordingly, we design four reward functions for rule-based RL training.

Accuracy Reward (R_{acc}). We reward the model based on soft exact match (EM), considering a response correct and $\text{EM}(a, a_{gt})$ equal to 1 if the normalized predicted answer a is a substring of the ground truth a_{gt}, or vice versa. To prevent the reward from becoming too sparse, we enhance this signal by including the recall metric. Formally,

$$R_{\text{acc}} = \frac{\mathbb{I}\left(\text{EM}(a, a_{gt}) = 1\right) + \text{Recall}(a, a_{gt})}{2}, \quad (3)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. By incorporating the recall metric, we prevent the wholesale rejection of semantically relevant yet not exact-match samples, while assigning higher rewards to perfectly matched outputs, thereby enhancing the model's generalization capability.

Stepwise Attribution Reward (R_{step}). To ensure each reasoning step is grounded in the associated evidence attribution, we design a stepwise reward that measures the semantic alignment between each step and the evidence indicated by the bounding box. For each textual reasoning r_t involving a reference element, we crop the raw document image of the i_t-th page according to the predicted bounding box B_t and encode both the cropped sub-image and r_t using a multimodal retriever ColQwen2 (Faysse et al. 2024), which enables efficient document indexing through visual features and handles dynamic resolutions. Formally,

$$e_{\text{img}}^{(t)} = \text{Norm}(\text{enc}_{\text{img}}(\textbf{crop}(B_t))), \quad (4)$$

$$e_{\text{txt}}^{(t)} = \text{Norm}(\text{enc}_{\text{txt}}(r_t)), \quad (5)$$

where $\text{Norm}(\cdot)$ denotes L_2 normalization for consistent cosine similarity evaluation. $\text{enc}_{\text{img}}(\cdot)$ and $\text{enc}_{\text{txt}}(\cdot)$ represent the image and text encoders of the retriever, respectively. **crop**$(\cdot)$ denotes the extraction of a sub-image from the original document page, delimited by the bounding box B_t. The stepwise attribution reward is defined as:

$$S = \min_{1 \le k \le K} \cos(e_{\text{img}}^{(k)}, e_{\text{txt}}^{(k)}), \quad (6)$$

$$I = \max_{1 \le i,j \le K,\, i \ne j} \text{IoU}(B_i, B_j), \quad (7)$$

$$R_{\text{step}} = \frac{\mathbb{I}(S \ge \tau) + \mathbb{I}(I \le \delta)}{2} \cdot \mathbb{I}(R_{\text{acc}} \ge \epsilon), \quad (8)$$

where K is the number of context-evidence pairs and $\cos(\cdot)$ denotes the cosine similarity measure. We require all pairs in the CoE reasoning process to exceed the similarity threshold τ to ensure proper alignment. Nevertheless, the model may exploit this mechanism by repeating bounding boxes to satisfy the reward, which undermines coarse-to-fine evidence attribution and rich visual cues.

To mitigate such a phenomenon, we introduce a constraint on bounding box overlap by computing the maximum pairwise intersection over union (IoU) I among bounding boxes from the reasoning steps (Equation 7) and enforcing $I \le \delta$ in Equation 8. This constraint promotes attribution diversity and encourages progressive grounding across reasoning steps. Furthermore, the reward is constrained by R_{acc}, ensuring that only faithful CoE processes contributing to correct answers are reinforced. This formulation guides the model to treat visual grounding as an internal retrieval problem, supporting stepwise and fine-grained visual attribution.

Grounding Reward (R_{ground}). The model needs to select the relevant source from the CoE reasoning process as evidence to support the final answer. We measure its precision by computing the IoU between the predicted bounding box (i^*, B_{ans}) and the ground-truth evidence (i_{gt}, B_{gt}):

$$R_{\text{ground}} = \mathbb{I}\left(\text{IoU}(B_{ans}, B_{gt}) > 0.5\right), \text{s.t. } i^* = i_{gt}. \quad (9)$$

By penalizing misaligned evidence attribution, R_{ground} drives the model to identify correct content within the source page i_{gt} instead of blindly collecting irrelevant information.

Format Reward (R_{format}). We prompt the model to conduct CoE reasoning on visual documents and generate answers. Well-formatted outputs contain CoE reasoning within `<think>...</think>` tags and answers with supporting evidence within `<answer>...</answer>` tags.

$$R_{\text{format}} = \begin{cases} 1, & \text{if the format is correct,} \\ -1, & \text{otherwise.} \end{cases} \quad (10)$$

By assigning negative rewards to incorrectly formatted outputs, we impose stricter output constraints and accelerate convergence toward the desired format, enabling more targeted optimization of the remaining reward components.

Training Algorithm. As shown in Figure 2, we sample 5% training data from the raw dataset, initialize the policy model π_θ from the cold-start checkpoint $\mathcal{M}_{\text{dist}}$, and adopt the GRPO (Shao et al. 2024b) algorithm, combined with the fine-grained reward design (Shen et al. 2025). GRPO, which supports rule-based rewards and optimizes the policy π_θ by sampling a group of candidate outputs for each query and computing a group-relative advantage, eliminates the need for training a critic model. We iteratively update π_θ, encouraging it to produce responses with higher relative advantages within the sampled group under the GRPO objective.

4 Experiment Setup

Datasets

We conducted the experiments on the VISA benchmark (Ma et al. 2024b), which is the first dataset designed for visual evidence attribution in real-world VD-RAG scenarios and comprises three subsets: (1) **Wiki-VISA** is derived from the Natural Questions (NQ) (Kwiatkowski et al. 2019). VISA renders the Wikipedia pages and identifies the HTML element containing the answer with a bounding box. (2) **Paper-VISA** builds upon PubLayNet (Zhong, Tang, and Yepes 2019). VISA synthesizes QA pairs grounded in annotated layouts. (3) **FineWeb-VISA** extends FineWeb-edu (Penedo et al. 2024) by selecting passages longer than 50 tokens and synthesizing grounded QA pairs. In the single-image setup, each query is paired with a source document page, a short answer extracted from it, and an evidence bounding box.

When extended to multiple images, for each query q, two additional document images are randomly sampled from the top K screenshots retrieved by Document Screenshot Embedding (DSE) (Ma et al. 2024a) within the VISA dataset, then merged with the source document page as input. For no-answer scenarios, the source document page is replaced with an irrelevant page from the dataset. Example cases are shown in Figure 5 in Appendix B.

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
Proprietary and Open-source Models, Direct Answer									
Gemini2.0 flash †	-	73.3	-	46.8	-	28.2	-	26.7	-
Qwen-VL max †	-	63.1	-	46.6	-	48.7	-	32.1	-
mPLUG-DocOwl2	8B	23.6	-	20.3	-	19.3	-	21.5	-
Qwen2.5-VL	7B	67.7	-	38.2	-	54.1	-	34.6	-
Qwen2.5-VL	32B	66.3	-	36.0	-	60.8	-	38.8	-
InternVL2.5	8B	45.9	-	42.4	-	37.7	-	32.1	-
LLaVA-OneVision	7B	48.4	-	28.7	-	38.5	-	30.8	-
LLaVA-CoT	11B	45.3	-	28.8	-	-	-	-	-
R1-OneVision	7B	58.0	-	24.4	-	57.2	-	31.4	-
Proprietary Model, Full Data Training, Direct Answer & Attribution									
VISA †	7B	74.1	28.3	49.6	62.2	58.7	31.6	54.1	65.6
Qwen2.5-VL (DA)	7B	69.4	0.80	43.8	2.87	52.5	0.97	37.2	5.56
LAT-Ind.	7B	**73.6**	53.7	45.4	**49.9**	64.5	38.0	**51.2**	46.9
LAT-Full	7B	73.1	**57.8**	**46.2**	48.4	**64.8**	**41.4**	50.6	**49.3**
Δ Vanilla model	-	+4.20	+57.0	+2.40	+47.0	+12.3	+40.4	+14.0	+43.7

Table 1: Performance comparison on Paper- and Wiki-VISA in both single- and multi-image settings. Bold indicates the best score in each column. † denotes the proprietary model and those fine-tuned on the full in-domain dataset, serving as an upper-bound baseline. "DA" refers to the zero-shot prompting setting for direct answer & attribution without training.

Baseline

To evaluate the effectiveness of LAT, we compared it against three categories of baselines: **(1) Proprietary and open-source models**, including Gemini (Anil et al. 2025), Qwen-VL Max (Bai et al. 2023), mPLUG-DocOwl2 (Hu et al. 2024), InternVL2.5 (Chen et al. 2025c), LLaVA-OneVision (Li et al. 2024), and Qwen2.5-VL (Bai et al. 2025a). **(2) Reasoning models**, including LLaVA-CoT (Xu et al. 2025), which is trained via SFT on CoT data, and R1-OneVision (Yang et al. 2025), which is optimized with RL in general-purpose scenarios. **(3) Attribution-supervised model:** VISA-7B (Ma et al. 2024b), trained to generate answers with direct attribution. These models serve as baselines to evaluate whether LAT balances answer accuracy with attribution precision and achieves CoE reasoning under limited supervision.

Training Details

We used Qwen2.5-VL-7B-Instruct (Bai et al. 2025a) as the backbone and applied LoRA (Hu et al. 2022) for parameter-efficient fine-tuning, with rank r=64 and scaling factor α=64. Following the pipeline in Figure 2, we performed SFT on $\mathcal{D}_{\text{final}}$ using a learning rate of 1e$-$4, followed by RL with 5e$-$5. During RL training, LAT was trained on 5% of QA pairs sampled from the raw dataset. After each stage, we merged the LoRA parameters for subsequent training.

In the multi-image setting, each query is paired with three retrieved documents provided in the `candidates` field, as included in the VISA dataset. When no relevant information is available, the model is trained to output "No answer". We initialized the multi-image model from the single-image trained version and further performed SFT using multi-image CoE data in $\mathcal{D}_{\text{final}}$, fine-tuning the LoRA adapter of the LM while keeping the vision transformer (ViT) frozen to reduce GPU memory usage. Additional experimental details are provided in Appendix A.

Evaluation

We measured the performance across three dimensions: answer accuracy, evidence grounding, and stepwise attribution quality. Specifically, we reported answer accuracy using **soft Exact Match (EM)**, and evaluated grounding precision by computing **IoU@0.5**, which measures the proportion of the predicted box B_{ans} whose IoU with the ground truth evidence exceeds 0.5. To evaluate the quality of stepwise visual attribution, we employed the **Stepwise Attribution (SA)** reward function. We adopted the default threshold τ=0.3 in Equation 8 to determine whether a step is correctly verified.

Following the evaluation of VISA (Ma et al. 2024b), we assessed the performance on both the Paper- and Wiki-VISA datasets under two settings. (1) **Single-image**: The model is provided solely with the source document and evaluated across three dimensions. (2) **Multi-image**: The model is additionally required to identify the source document from a set of retrieved candidates. For reproducibility, we adopt greedy decoding as the decoding strategy during evaluation.

5 Results and Analysis

Main Results

Attribution-aware performance under in-domain and cross-domain settings. To assess both answer accuracy and attribution precision, we evaluated models under direct answer and attribution-aware settings. We define an in-domain setup as evaluation on datasets from the same distribution as the training data. As shown in Table 1, LAT trained on in-domain data (**LAT-Ind.**) outperforms open-source models. Compared to the vanilla model, LAT enhances both answer correctness (EM, **+7.95%**) and evidence

Models	Param.	Wiki-VISA (Single)			Paper-VISA (Single)			Wiki-VISA (Multi)			Paper-VISA (Multi)		
	Size	EM	IoU@0.5	SA	EM	IoU@0.5	SA	EM	IoU@0.5	SA	EM	IoU@0.5	SA
Proprietary Models, CoE Reasoning													
Gemini2.0 flash †	-	52.0	4.47	24.0	46.4	4.72	14.1	32.1	1.20	5.20	27.0	3.24	20.0
Qwen-VL max †	-	55.2	0.17	10.3	47.8	2.87	8.6	35.8	0.10	5.90	32.6	1.85	12.5
Open-source Models, CoE Reasoning													
Qwen2.5-VL	7B	60.4	2.20	13.9	36.9	3.80	12.4	54.9	11.0	12.5	39.6	16.2	29.5
+one shot ICL	7B	61.8	1.37	12.2	37.3	8.30	29.3	54.2	9.20	11.9	39.5	18.7	30.9
Qwen2.5-VL	32B	62.8	9.27	0.16	35.2	2.50	0.11	62.5	18.7	0.60	43.3	19.0	0.40
+one shot ICL	32B	61.7	9.13	5.23	35.0	4.12	1.25	64.1	22.1	1.87	42.6	19.6	0.51
LAT-Ind.	7B	**73.6**	53.7	**64.6**	45.4	**49.9**	**35.5**	64.5	38.0	71.8	**51.2**	46.9	46.3
LAT-Full	7B	73.1	**57.8**	59.6	**46.2**	48.4	33.8	**64.8**	**41.4**	**75.2**	50.6	**49.3**	**53.2**
Δ Vanilla model	-	+13.2	+55.6	+50.7	+9.3	+46.1	+23.1	+9.9	+30.4	+62.7	+11.6	+33.1	+23.7

Table 2: Performance comparison for Chain-of-Evidence (CoE) reasoning processes on Paper- and Wiki-VISA datasets. Bold indicates the best score in each column. Results include both result accuracy (EM, IoU@0.5) and process quality (SA) metrics.

grounding (IoU@0.5, **+44.6%**) by optimizing attribution-aware reasoning through CoE-guided RL. The improvements are evident in multi-page scenarios, where the complexity of evidence selection emphasizes the advantages of our approach. In particular, for unanswerable cases, LAT outputs "No answer" and achieves an average precision of 65%, highlighting its robustness in handling such scenarios.

Table 2 highlights the effectiveness of LAT in improving CoE reasoning quality. To assess the impact of in-context learning (ICL) (Brown et al. 2020), we included an annotated CoE example as a prompt. While this yields a moderate improvement (SA, +4.0%), suggesting that demonstrations can partially guide the generation of structured reasoning, it has a limited effect on the result. LAT achieves a substantial SA gain of **37.5%**, demonstrating its ability to ground each reasoning step accurately. Moreover, LAT maintains high answer accuracy, indicating an alignment between faithful reasoning and correct outcomes. To assess generalization across domains, we trained **LAT-Full** on the sampled subset from all datasets. Compared to the in-domain variant, **LAT-Full** shows further improvements (e.g., Wiki-VISA Multi, IoU@0.5: 38.0%→**41.4%**; SA: 71.8%→**75.2%**), exhibiting generalization across diverse document distributions.

CoE reasoning performance with limited supervision. Unlike VISA, which relies on large-scale (100k) supervised data and directly links the final answer to supporting evidence without reasoning, LAT is trained on only 5% of raw QA pairs during the RL stage. In low-resource settings, LAT achieves comparable performance to VISA-7B in answer accuracy and attribution precision, while maintaining traceable CoE reasoning. Notably, on high-resolution Wiki-VISA images, LAT outperforms VISA-7B, demonstrating robustness in visually complex scenarios under limited supervision.

To ensure fair comparison, we established an SFT baseline by training the model with VISA's supervision protocol on the same data subset and experimental setup used for our approach. Figure 3a demonstrates LAT's superior performance in both answer accuracy and evidence precision, with greater improvements observed in the multi-image setting (Figure 3b). Meanwhile, Table 3 highlights LAT's generalization across datasets, outperforming SFT by 1.7% in EM and 6.2% in IoU@0.5 in the "Paper→Wiki" transfer setting, while preserving the vanilla model's performance on Wiki- (EM: 67.7%) and Paper-VISA (EM: 38.2%). This demonstrates that LAT improves transfer effectiveness without sacrificing adaptability to diverse data types.

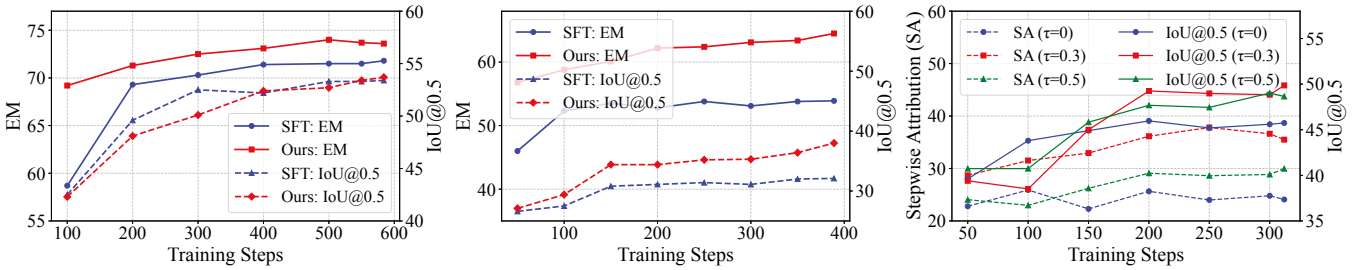
Train→Eval	Method	EM	IoU@0.5
Paper→Wiki (Single)	SFT	66.0	29.4
	LAT-Ind.	$67.7_{\uparrow 1.7}$	$35.6_{\uparrow 6.2}$
Paper→Wiki (Multi)	SFT	48.7	10.3
	LAT-Ind.	$57.3_{\uparrow 8.6}$	$21.4_{\uparrow 11.1}$

Table 3: LAT demonstrates robust generalization with cross-domain transfer between Paper-VISA and Wiki-VISA.

Ablation Study

Effectiveness of reward components. We conducted ablation studies on both datasets to examine the contributions of individual components in our reward formulation (Table 4). We first analyzed the impact of distillation based on annotated CoE reasoning trajectories. The model $\mathcal{M}_{\text{dist}}$, obtained through fine-tuning during the cold start stage, shows an average improvement of **6.48%** in EM and **20.9%** in IoU@0.5 compared to the vanilla model. This indicates that distillation improves adherence to CoE reasoning formats.

Next, we assess the impact of stepwise attribution by ablating the process reward R_{step}. The model fails to align intermediate reasoning steps with visual evidence, resulting in a 15.5% reduction in SA. Since the evidence grounding of the answer is inherently linked to the quality of intermediate visual attribution ($B_{ans} \in \mathcal{B}$), we also observed a decline in IoU@0.5, confirming the necessity of step-level attribution. Meanwhile, without the overlap constraint $I \leq \delta$ in R_{step}, the model tends to reuse large and redundant regions across reasoning steps. This behavior undermines the coarse-to-fine grounding strategy and reduces attribution fidelity.



(a) LAT vs. SFT. (Wiki-VISA Single) (b) LAT vs. SFT. (Wiki-VISA Multi) (c) Performance variation with different τ.

Figure 3: Comparison of LAT and SFT performance across different settings and an ablation study on threshold τ.

Models	Wiki-VISA (Single)			Paper-VISA (Single)			Wiki-VISA (Multi)			Paper-VISA (Multi)		
	EM	IoU@0.5	SA	EM	IoU@0.5	SA	EM	IoU@0.5	SA	EM	IoU@0.5	SA
Vanilla model (Qwen2.5-VL-7B)	60.4	2.20	13.9	36.9	3.80	12.4	54.9	11.0	12.5	39.6	16.2	29.5
$\mathcal{M}_{\text{dist}}$ (Stage I)	67.1	25.6	32.8	44.7	27.0	13.0	59.4	27.8	44.0	46.5	36.3	28.2
LAT-Ind. (Stage II)	**73.6**	**53.7**	**64.6**	**45.4**	**49.9**	35.5	**64.5**	**38.0**	71.8	**51.2**	**46.9**	46.3
w/o. R_{step}	73.1	49.8	43.8	45.0	45.7	24.1	61.8	35.0	48.0	49.0	44.8	40.5
w/o. $R_{\text{acc}}, R_{\text{ground}}$	72.7	30.4	59.8	44.9	28.7	44.2	59.0	31.8	67.1	50.7	41.3	54.7
w/o. $R_{\text{acc}}, R_{\text{ground}}, (R_{\text{acc}} \geq \epsilon)$	67.0	28.9	56.0	45.0	22.6	**60.7**	31.7	22.5	**81.7**	37.6	30.3	**63.2**

Table 4: Ablation study results on LAT. "w/o." denotes the results of models trained without the corresponding reward function.

Aligning reasoning and result objectives. We further analyzed the role of jointly optimizing process- and result-level rewards, where we removed the supervision from the final answer (w/o. $R_{\text{acc}}, R_{\text{ground}}$) and used it solely as a filter for valid reasoning paths. The model maintains moderate EM performance, indicating that consistency between the reasoning process and the answer inherently provides useful training signals. However, when supervision $R_{\text{acc}} \geq \epsilon$ is removed from R_{step}, the training objective is reduced to aligning only the reasoning format, resulting in significant degradation in both answer accuracy and evidence grounding precision. This suggests that R_{step} and result-level rewards work synergistically, where process rewards guide reasoning coherence while answer supervision ensures factual accuracy and proper grounding to source content.

Further Analysis

Sensitivity analysis of attribution threshold τ. To evaluate the alignment quality between visual evidence and textual references in CoE reasoning, we introduced a similarity threshold τ in the stepwise attribution reward function to distinguish positive from negative samples. Based on the synthetic CoE dataset $\mathcal{D}_{\text{final}}$, we computed semantic similarity scores across different answer types. As shown in Figure 8 in Appendix E, we reported the range of similarity values for each category, excluding pairs with zero scores. We set the default threshold to 0.3 based on the distribution analysis.

Given the parameter sensitivity, we further conducted experiments on both datasets to evaluate the robustness of τ across different document types. Specifically, we compared a high-threshold setting (τ=0.5), a no-step variant (equivalent to τ=0 or 1, where all steps are uniformly rewarded), and the default setting. As shown in Figure 3c and Figure 4 (Appendix E), τ=0.3 achieves a better balance, yielding consistent improvements in both IoU@0.5 and SA throughout training. In contrast, the high threshold fails to sustain performance gains, as such strict criteria make it difficult to sample sufficient positive instances, while the no-step variant underperforms due to the lack of fine-grained attribution guidance. These results suggest that attribution supervision at τ=0.3 offers relatively effective guidance for stepwise grounding. Additional analyses are reported in Appendix E.

Traceable Reasoning with the CoE Paradigm. The CoE paradigm achieves stepwise visual attribution, generating a traceable reasoning process toward the final answer. Through the LAT framework, we leverage stepwise rewards to achieve an average SA of 57.1%. Meanwhile, by penalizing repetitive reasoning processes in Equation 7, we encourage the model to generate diverse and fine-grained reasoning. As illustrated in Figure 12–17 in Appendix F, LAT accurately identifies the answer regions and generates faithful reasoning paths that closely align with the visual evidence.

6 Conclusion

In this paper, we introduce CoE, a reasoning paradigm that unifies CoT with stepwise visual evidence attribution. To achieve CoE, we propose LAT, a reinforcement learning framework that aligns the intermediate process to mitigate ungrounded reasoning for the visual evidence attribution task in VD-RAG. By incorporating stepwise rewards under the GRPO algorithm, LAT facilitates verification at each reasoning step. Experiments on Paper- and Wiki-VISA show that LAT outperforms baselines. We hope this work inspires further research on enhancing the verifiability of VD-RAG.

Acknowledgements

This work was supported in part by the grants from National Science and Technology Major Project (No. 2023ZD0121104), and National Natural Science Foundation of China (No.62222213, 62072423).

References

Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025a. Qwen2.5-VL Technical Report. arXiv:2502.13923.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2025b. Hallucination of Multimodal Large Language Models: A Survey. arXiv:2404.18930.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Cao, M.; Zhao, H.; Zhang, C.; Chang, X.; Reid, I.; and Liang, X. 2025. Ground-R1: Incentivizing Grounded Visual Reasoning via Reinforcement Learning. arXiv:2505.20272.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. arXiv:2306.15195.

Chen, Y.; Liu, S.; Lyu, Y.; Zhang, C.; Shi, J.; and Xu, T. 2025a. Xiangqi-R1: Enhancing Spatial Strategic Reasoning in LLMs for Chinese Chess via Reinforcement Learning. arXiv:2507.12215.

Chen, Y.; Lyu, Y.; Liu, S.; Zhang, C.; Lv, J.; and Xu, T. 2025b. Think Wider, Detect Sharper: Reinforced Reference Coverage for Document-Level Self-Contradiction Detection. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 1273–1288. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-332-6.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; Gu, L.; Wang, X.; Li, Q.; Ren, Y.; Chen, Z.; Luo, J.; Wang, J.; Jiang, T.; Wang, B.; He, C.; Shi, B.; Zhang, X.; Lv, H.; Wang, Y.; Shao, W.; Chu, P.; Tu, Z.; He, T.; Wu, Z.; Deng, H.; Ge, J.; Chen, K.; Zhang, K.; Wang, L.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2025c. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. arXiv:2412.05271.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. arXiv:2501.17161.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; et al. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. arXiv:2507.06261.

Faysse, M.; Sibille, H.; Wu, T.; Omrani, B.; Viaud, G.; Hudelot, C.; and Colombo, P. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Gao, T.; Yen, H.; Yu, J.; and Chen, D. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Hu, A.; Xu, H.; Zhang, L.; Ye, J.; Yan, M.; Zhang, J.; Jin, Q.; Huang, F.; and Zhou, J. 2024. mPLUG-DocOwl2: High-resolution Compressing for OCR-free Multi-page Document Understanding. arXiv:2409.03420.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2024. LLaVA-OneVision: Easy Visual Task Transfer. arXiv:2408.03326.

Li, Z.; Luo, R.; Zhang, J.; Qiu, M.; Huang, X.-J.; and Wei, Z. 2025. VoCoT: Unleashing Visually Grounded Multi-Step Reasoning in Large Multi-Modal Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3769–3798.

Ma, X.; Lin, S.-C.; Li, M.; Chen, W.; and Lin, J. 2024a. Unifying Multimodal Retrieval via Document Screenshot Embedding. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6492–6505. Miami, Florida, USA: Association for Computational Linguistics.

Ma, X.; Zhuang, S.; Koopman, B.; Zuccon, G.; Chen, W.; and Lin, J. 2024b. VISA: Retrieval Augmented Generation with Visual Source Attribution. arXiv:2412.14457.

Ni, M.; Yang, Z.; Li, L.; Lin, C.-C.; Lin, K.; Zuo, W.; and Wang, L. 2025. Point-RFT: Improving Multimodal Reasoning with Visually Grounded Reinforcement Finetuning. arXiv:2505.19702.

Penedo, G.; Kydlíček, H.; Lozhkov, A.; Mitchell, M.; Raffel, C. A.; Von Werra, L.; Wolf, T.; et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37: 30811–30849.

Peng, Y.; Zhang, G.; Zhang, M.; You, Z.; Liu, J.; Zhu, Q.; Yang, K.; Xu, X.; Geng, X.; and Yang, X. 2025. LMM-R1: Empowering 3B LMMs with Strong Reasoning Abilities Through Two-Stage Rule-Based RL. arXiv:2503.07536.

Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; and Wei, F. 2023. Kosmos-2: Grounding Multimodal Large Language Models to the World. arXiv:2306.14824.

Qi, J.; Ding, M.; Wang, W.; Bai, Y.; Lv, Q.; Hong, W.; Xu, B.; Hou, L.; Li, J.; Dong, Y.; and Tang, J. 2025. CogCoM: A Visual Language Model with Chain-of-Manipulations Reasoning. arXiv:2402.04236.

Shao, H.; Qian, S.; Xiao, H.; Song, G.; Zong, Z.; Wang, L.; Liu, Y.; and Li, H. 2024a. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37: 8612–8642.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y. K.; Wu, Y.; and Guo, D. 2024b. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. arXiv:2402.03300.

Shen, H.; Liu, P.; Li, J.; Fang, C.; Ma, Y.; Liao, J.; Shen, Q.; Zhang, Z.; Zhao, K.; Zhang, Q.; Xu, R.; and Zhao, T. 2025. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. arXiv:2504.07615.

Wang, J.; Kang, Z.; Wang, H.; Jiang, H.; Li, J.; Wu, B.; Wang, Y.; Ran, J.; Liang, X.; Feng, C.; and Xiao, J. 2025. VGR: Visual Grounded Reasoning. arXiv:2506.11991.

Wu, Q.; Yang, X.; Zhou, Y.; Fang, C.; Song, B.; Sun, X.; and Ji, R. 2025. Grounded Chain-of-Thought for Multimodal Large Language Models. arXiv:2503.12799.

Xia, J.; Tong, B.; Zang, Y.; Shao, R.; and Zhou, K. 2025. Bootstrapping Grounded Chain-of-Thought in Multimodal LLMs for Data-Efficient Model Adaptation. arXiv:2507.02859.

Xu, G.; Jin, P.; Wu, Z.; Li, H.; Song, Y.; Sun, L.; and Yuan, L. 2025. LLaVA-CoT: Let Vision Language Models Reason Step-by-Step. arXiv:2411.10440.

Yang, Y.; He, X.; Pan, H.; Jiang, X.; Deng, Y.; Yang, X.; Lu, H.; Yin, D.; Rao, F.; Zhu, M.; Zhang, B.; and Chen, W. 2025. R1-Onevision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization. arXiv:2503.10615.

Ye, X.; Sun, R.; Arik, S.; and Pfister, T. 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)*, 6237–6251.

Zhang, H.; Shen, S.; Xu, B.; Huang, Z.; Wu, J.; Sha, J.; and Wang, S. 2024. Item-difficulty-aware learning path recommendation: From a real walking perspective. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4167–4178.

Zhong, X.; Tang, J.; and Yepes, A. J. 2019. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, 1015–1022. IEEE.

Zhu, J.; Liu, S.; Yu, Y.; Tang, B.; Yan, Y.; Li, Z.; Xiong, F.; Xu, T.; and Blaschko, M. B. 2024. FastMem: Fast Memorization of Prompt Improves Context Awareness of Large Language Models. arXiv:2406.16069.