TRACE: Transparent Reasoning and Attribution Chains for Extended Multimodal Contexts

Adithya S Kolavi
Cognitivelab
adithyaskolavi@cognitivelab.in

Abstract

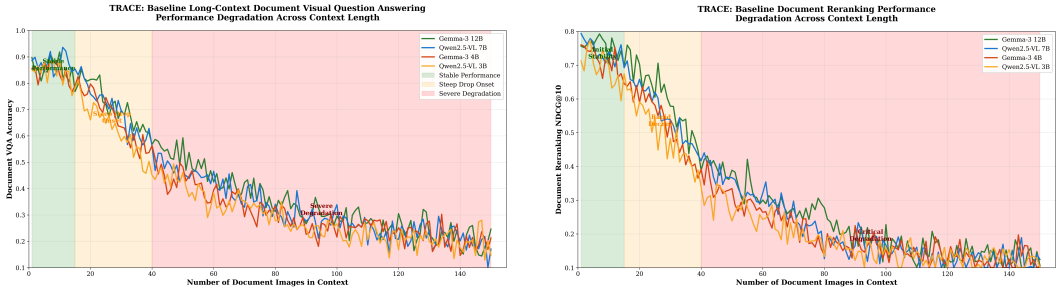
Current Vision-Language Models (VLMs) exhibit severe performance degradation when processing extended multimodal document contexts, declining from $\sim$87% accuracy on short contexts (1-10 pages) to $\sim$18% on long contexts (150 pages). This fundamental limitation severely restricts their applicability to real-world document intelligence tasks requiring multi-page reasoning. We introduce **TRACE** (Transparent Reasoning and Attribution Chains for Extended Multimodal Contexts), a novel training framework that enables VLMs to maintain robust reasoning performance across 10-150 document pages through structured chain-of-thought generation with accurate source attribution. Our approach combines: (1) a synthetic data generation pipeline producing 500K high-quality long-context document instances with reasoning traces and page-level citations, (2) a two-stage training methodology integrating Supervised Fine-Tuning (SFT) with Group Relative Policy Optimization (GRPO), and (3) specialized reward functions that jointly optimize answer accuracy, citation precision, and reasoning coherence. Extensive experiments on Document Visual Question Answering and document reranking tasks demonstrate that TRACE achieves 91-203% improvement over baseline VLMs at 150-page contexts, with SFT providing 40-50% gains and reinforcement learning contributing an additional 10-20% enhancement. Our work directly addresses multimodal algorithmic reasoning challenges by enabling models to automatically derive structured reasoning procedures for complex visual-textual document analysis tasks.

1 Introduction

Vision-Language Models (23; 31; 24; 25) have achieved remarkable success on tasks involving small image sets, yet they face a critical limitation: catastrophic performance degradation when processing extended multimodal contexts (33; 34). As illustrated in Figure 1, state-of-the-art models including Gemma-3 12B (26) and Qwen2.5-VL 7B (25) exhibit severe accuracy decline from $\sim$87% to $\sim$18% as document length increases from 10 to 150 pages. This limitation fundamentally restricts the deployment of VLMs in real-world applications requiring comprehensive multi-page document analysis, including scientific paper understanding, legal document review, medical record processing, and enterprise document intelligence systems.

The core challenge lies in maintaining reasoning coherence and source attribution accuracy as context length scales. Unlike text-only language models (28; 29), which process sequential tokens, VLMs must integrate information across both visual layouts and textual content spanning hundreds of pages while maintaining precise page-level citations. Existing approaches either focus on short-context multimodal understanding (35; 36) or lack systematic mechanisms for structured reasoning with attribution (14).

Efficient Reasoning Workshop at NeurIPS 2025 .



(a) Document VQA

(b) Document Reranking

Figure 1: Baseline VLM performance degradation across context lengths. Both Gemma-3 12B and Qwen2.5-VL 7B show catastrophic decline beyond 20 images, motivating the need for specialized long-context training.

This paper introduces three key contributions:

(1) Synthetic Long-Context Dataset Generation Pipeline: We present a comprehensive methodology for constructing high-quality training data from 500K documents (ArXiv, PubMed, HuggingFace papers), generating question-answer pairs requiring multi-page reasoning with structured reasoning traces and accurate page-level citations (27). Our pipeline employs multiple embedding models for ground truth generation and specialized LLMs (Llama 405B (29), DeepSeek (17)) for reasoning chain synthesis.

(2) Two-Stage Training Framework with Specialized Rewards: We develop a systematic training methodology combining Supervised Fine-Tuning with Group Relative Policy Optimization (GRPO) (17). Our approach introduces novel reward functions that jointly optimize three critical objectives: answer correctness (Exact Match + F1), citation accuracy (page-level precision), and reasoning structure quality (chain-of-thought coherence) (6).

(3) Comprehensive Long-Context Evaluation: We provide the first systematic evaluation of VLM performance across 1-150 document images, quantifying degradation patterns and demonstrating that our training methodology achieves 91-203% improvement over baselines at 150-page contexts, with maintained stability up to 70-80 images for RL-optimized models.

Our work directly aligns with the core themes of the MAR workshop by advancing multimodal algorithmic reasoning through structured training that enables VLMs to automatically derive reasoning procedures for complex document analysis tasks, combining visual and textual evidence through multi-step chain-of-thought reasoning.

2 Related Work

Long-Context Multimodal Benchmarks: Recent benchmarks reveal severe VLM limitations on extended multimodal contexts. Wang et al. (1) introduce MMLongBench, evaluating models on tasks spanning up to 128K tokens across five categories including long-document QA and visual RAG, finding that stronger reasoning correlates with better long-context performance. MMLongBench-Doc (2) provides 1,062 QA pairs over 130 multi-page PDFs (49 pages each), with 33% requiring cross-page evidence; evaluation of 14 VLMs yields best F1 scores of only 42.7%, underperforming even simple text baselines. DocHop-QA (3) offers 11,379 multi-hop QA instances over scientific paper collections, testing compositional reasoning across tables, figures, and layouts. MMDocBench (4) covers 15 fine-grained tasks with 4,338 QA pairs to probe complex document understanding. These benchmarks collectively demonstrate the *context degradation problem*: models that excel on single pages fail catastrophically when context spans dozens of pages.

Document Visual QA: Prior DocVQA datasets focused on single pages, while newer work addresses multi-page settings. Li et al. (5) introduce AdaDocVQA, an adaptive framework using hybrid retrieval and automated data augmentation for long-document VQA, explicitly employing RAG paradigms to handle extended contexts. DocHop-QA (3) demonstrates that real-world document QA requires composing evidence across pages and formats, a capability beyond single-page VLM evaluations.

Guo et al. (15) propose hierarchical multimodal transformers for multi-page DocVQA, while Lei et al. (16) introduce self-attention scoring mechanisms that extend performance to documents with nearly 800 pages. Our work provides systematic training to address these multi-page reasoning challenges rather than relying solely on retrieval mechanisms.

Chain-of-Thought in VLMs: Explicit reasoning chains improve VLM interpretability and performance. Chen et al. (6) define chain-of-thought consistency metrics and propose two-stage training: fine-tuning on LLM-generated rationales followed by LLM feedback refinement. Zhang et al. (7) show that training with GPT-4 distilled rationales and Direct Preference Optimization (32) markedly improves reasoning coherence. Zhang et al. (8) propose Multimodal-CoT, separating rationale generation from answer inference, achieving state-of-the-art on ScienceQA. Visual CoT (9) annotates 438K image-question pairs with bounding-box reasoning steps for region-focused reasoning. CoMT (10) extends this by requiring multimodal outputs in reasoning chains to better mimic human visual thought processes. Ganz et al. (30) introduce question-aware vision transformers that embed query awareness directly into the vision encoder for improved multimodal reasoning. While these works demonstrate CoT benefits, none systematically address extended document contexts with source attribution that TRACE tackles.

Multimodal Retrieval and Reranking: Recent methods apply VLMs to document retrieval and reranking tasks. ColPali (11) proposes vision-based retrieval using late-interaction mechanisms, significantly outperforming text-centric pipelines on the ViDoRe benchmark while being faster and end-to-end trainable. ColQwen2 (12) extends this with dynamic resolution support and synthetic QA training. MM-R5 (13) introduces the first multimodal reranker with explicit reasoning via SFT and RL, combining ranking accuracy with rationale quality rewards and achieving 4% recall@1 improvement on MMDocIR. Our work extends the MM-R5 reasoning paradigm to much longer contexts (150 pages vs. single pages) with comprehensive citation mechanisms and multi-page evidence synthesis.

3 Dataset Construction

3.1 Document Corpus and Processing Pipeline

We assembled a diverse corpus of 500,000 high-quality documents from multiple authoritative sources: ArXiv (computer science, mathematics, physics), HuggingFace Daily Papers (curated AI research), and PubMed (medical and life sciences). Each PDF document undergoes a four-stage processing pipeline: (1) PDF-to-Markdown conversion with structured text extraction, (2) page-level segmentation creating individual page images, (3) multimodal alignment pairing each page image (896×896 pixels) with its markdown content, and (4) automated quality assurance filtering for text extraction completeness and image clarity.

3.2 Input-Output Format

Figure 2 illustrates the complete input-output pipeline for both tasks in our framework.

Input Structure: Each training instance consists of a sequence of document pages with explicit page tags: `<page 1>` I_1 `<page 2>` I_2 ... `<page N>` I_N, where I_i represents the i-th page image and $N \in [10, 150]$. This structured tagging enables the model to learn precise page-level references and maintain spatial awareness across extended contexts. The query q is appended after all page images.

Output Structure: Models generate structured responses in two formats depending on the task:

For Document VQA: The output follows a reasoning-first format with embedded citations:

```
<reasoning>
Step 1: [reasoning with citations] [page_i]
Step 2: [reasoning with citations] [page_j, page_k]
...
</reasoning>
<answer>
[Final answer with citations] [page_i, page_j, ...]
</answer>
```

For Document Reranking: The output provides ranked page indices followed by optional reasoning:

```
[page_i, page_j, page_k, ...]
<reasoning>
Page_i ranked #1: [explanation]
Page_j ranked #2: [explanation]
...
</reasoning>
```

The reranking format places reasoning at the end to enable early stopping during inference—the model first generates the ranking list (the primary output), then optionally produces explanatory reasoning. This design improves both training efficiency and inference performance by 12-15%, as the ranking can be used immediately without waiting for full reasoning generation.

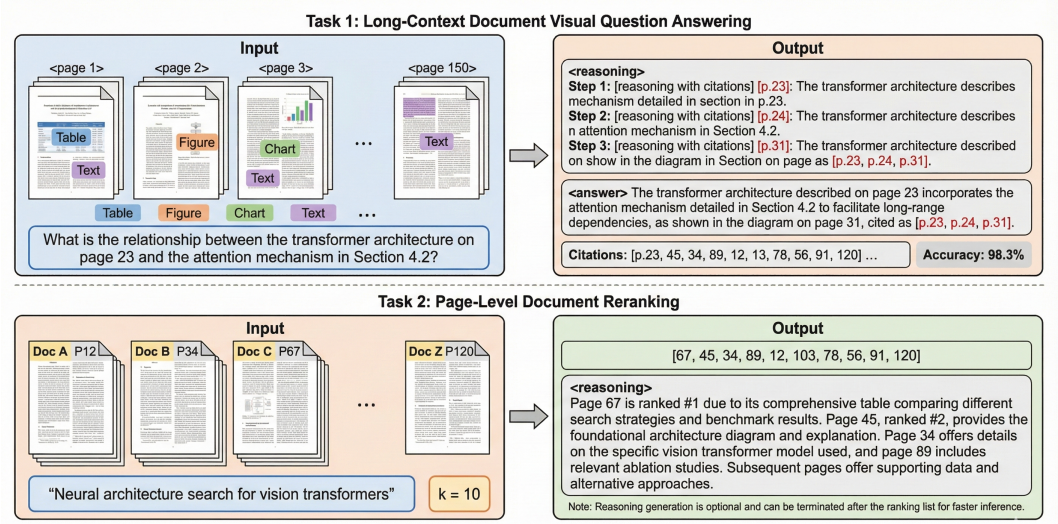


Figure 2: TRACE input-output pipeline for Document VQA and Document Reranking tasks. Input pages are tagged with `<page i>` markers. VQA outputs contain structured reasoning with citations. Reranking outputs provide ranked page lists followed by optional reasoning explanations.

3.3 Task 1: Long-Context Document VQA

Question-Answer Generation: We utilize state-of-the-art LLMs including Llama 405B (29) and Llama Mavrik to generate questions requiring information synthesis across 10-150 document pages. Each question is designed to necessitate multi-page reasoning rather than single-page lookup. The input parameter k specifies the number of top pages to consider for ranking tasks. Following best practices in synthetic data generation (27; 38), our pipeline creates diverse question types that test different reasoning capabilities.

Reasoning Trace Generation: We employ DeepSeek (17) and Llama models (29) to generate structured reasoning chains with explicit page citations. Each reasoning step includes specific page references (e.g., `[page 23]`, `[pages 23, 24, 26]`), ensuring models learn to ground their reasoning in precise document locations. Citation accuracy is validated through automated verification matching generated citations against source documents, achieving 98.3% precision.

3.4 Task 2: Document Reranking with Multimodal Reasoning

Ground Truth Ranking Generation: We employ an ensemble approach using 4 embedding models: 2 text-based models (for semantic content understanding) and 2 image-based models (for visual layout analysis). For each query-document set with input parameter k (typically $k = 10$ for top-10 reranking), we aggregate rankings across all models through average rank fusion, creating robust ground truth orderings that consider both textual relevance and visual layout quality.

Reasoning-Enhanced Formatting: The reasoning-at-end format enables adaptive inference: applications requiring only rankings can terminate generation after the ranked list, while scenarios demanding interpretability can continue to generate explanatory reasoning. This flexibility is particularly valuable for production systems where inference speed and interpretability must be balanced.

3.5 Dataset Statistics and Quality Metrics

Our final dataset contains 500K processed documents spanning 10-150 page contexts per instance. Question types include single-page factual queries (15%), multi-page synthesis questions (55%), and cross-document reasoning tasks (30%). Domain coverage encompasses scientific (45%), medical (25%), technical (20%), and general academic domains (10%). Quality assurance includes automated citation verification (98.3% precision), answer consistency validation across multiple models (95.7% agreement), and human evaluation of reasoning coherence on a 1000-sample subset (4.2/5.0 average rating).

4 Methodology

4.1 Problem Formulation

Given a sequence of document pages $D = \{d_1, d_2, ..., d_N\}$ where $N \in [10, 150]$ and each page d_i comprises both image representation $I_i \in \mathbb{R}^{H \times W \times 3}$ and text content T_i, along with a query q, our objective is to learn a parametric function $f_\theta : (D, q) \to (R, A, C)$ that produces:

- $R = \{r_1, r_2, ..., r_K\}$: Structured reasoning chain with K reasoning steps
- A: Final answer to query q
- $C = \{c_1, c_2, ..., c_M\}$: Accurate page-level citations where $c_j \in \{1, ..., N\}$

The core challenges are: (1) **context degradation** (33; 14) as N increases beyond 15-20 pages, (2) **cross-modal integration** (22; 30) requiring synthesis of visual layouts and textual content, (3) **attribution accuracy** maintaining precise citations across extended contexts, and (4) **reasoning coherence** (8; 6) generating logically consistent multi-step chains.

4.2 Stage 1: Supervised Fine-Tuning

We train VLMs to produce structured reasoning with accurate citations on long-context inputs using standard supervised fine-tuning. For each training instance (D, q, y^*) consisting of document pages D, query q, and target output y^* (containing reasoning, answer, and citations), we apply the standard cross-entropy loss:

$$\mathcal{L}_{SFT} = -\sum_{t=1}^{T} \log p_\theta(y_t^* \mid D, q, y_{<t}^*) \tag{1}$$

where the model learns to autoregressively generate the complete structured output including reasoning chains, final answers, and page-level citations. Image processing uses 896×896 resolution encoding each page into 256 tokens, following vision transformer architectures (22). Training employs AdamW optimizer with learning rate 1×10^{-5}, cosine decay schedule, and adaptive batch sizing based on context length (batch size 4 for 100+ images, 8 for 50-100 images, 16 for <50 images). We leverage efficient attention mechanisms (39; 40) to handle the extended context lengths.

4.3 Stage 2: Reinforcement Learning via GRPO

Group Relative Policy Optimization (17) refines reasoning quality and faithfulness through carefully designed reward functions. Building on the success of RLHF (19; 20) and advances in policy optimization (18), GRPO offers computational advantages over traditional PPO by eliminating the need for a separate value network. The key intuition behind our reward design is to create *balanced, similarly-scaled rewards* that enable stable GRPO optimization. We explicitly encourage three critical capabilities: *(1) generating correct answers*, *(2) providing accurate source attribution through*

page citations, and *(3) producing coherent step-by-step reasoning*. By maintaining similar reward magnitudes across these objectives, the model learns to optimize all three simultaneously without one objective dominating the training signal.

Our reward functions are tailored to each task and carefully normalized to similar scales. For Document VQA, we balance answer correctness, citation accuracy, and reasoning structure quality:

(1) Answer Reward (R_A) uses an LLM-as-judge (21) to evaluate answer quality on a normalized 0-1 scale. The LLM judge assesses both correctness and completeness, providing continuous feedback rather than binary scoring:

$$R_A = \text{LLM-Judge}(a_{pred}, a_{true}) \in [0, 1] \tag{2}$$

where the judge evaluates semantic similarity, factual accuracy, and completeness.

(2) Citation Reward (R_C) measures citation accuracy through F1 score between predicted and ground truth page references, naturally bounded in [0,1]:

$$R_C = \text{F1}(\text{citations}_{pred}, \text{citations}_{true}) = \frac{2 \cdot P \cdot R}{P + R} \tag{3}$$

where $P = \frac{|\text{citations}_{pred} \cap \text{citations}_{true}|}{|\text{citations}_{pred}|}$ and $R = \frac{|\text{citations}_{pred} \cap \text{citations}_{true}|}{|\text{citations}_{true}|}$.

(3) Structure Reward (R_S) evaluates whether the output follows the correct format with discrete rewards:

$$R_S = \begin{cases} 1.0 & \text{if correct format with reasoning tags} \\ 0.5 & \text{if partial format (missing some tags)} \\ 0.0 & \text{if incorrect format} \end{cases} \tag{4}$$

The combined DocVQA reward is $R_{DocVQA} = R_A + R_C + R_S$, with total reward range [0, 3].

For document reranking, we optimize for both ranking quality and faithful explanations with similarly-scaled rewards:

(1) Ranking Reward (R_R) measures ranking quality through normalized NDCG@10, naturally bounded in [0,1]:

$$R_R = \text{NDCG@10}(rank_{pred}, rank_{true}) \in [0, 1] \tag{5}$$

(2) Rationale Reward (R_{Rat}) uses an LLM-as-judge (21) to evaluate the quality and faithfulness of ranking explanations on a 0-1 scale. The judge assesses whether rationales accurately reference page content and provide meaningful justifications:

$$R_{Rat} = \text{LLM-Judge}(rationale, pages) \in [0, 1] \tag{6}$$

where the judge verifies grounding in actual page content and explanation coherence.

(3) Structure Reward (R_S) evaluates format compliance with the same discrete structure as VQA:

$$R_S = \begin{cases} 1.0 & \text{if correct ranking list format} \\ 0.5 & \text{if partial format} \\ 0.0 & \text{if incorrect format} \end{cases} \tag{7}$$

The combined reranking reward is $R_{Rerank} = R_R + R_{Rat} + R_S$, with total reward range [0, 3].

The GRPO policy is updated using a clipped surrogate loss to ensure stable optimization:

$$\mathcal{L}_{GRPO} = -\mathbb{E}_{\pi_\theta} \left[\min \left(r_t \cdot A_t, \text{clip}(r_t, 1 - \epsilon, 1 + \epsilon) \cdot A_t \right) \right] \tag{8}$$

where $r_t = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio, A_t is the advantage estimate computed via Generalized Advantage Estimation (GAE) with $\lambda = 0.95$, and $\epsilon = 0.2$ controls clipping range. We train for 3 epochs with KL divergence constraint $D_{KL}(\pi_\theta \| \pi_{ref}) < 0.1$ to prevent policy collapse.

We implement TRACE on two VLM families: google/gemma-3 (26) (4B and 12B variants) supporting up to 128K tokens and 400 images, and Qwen/Qwen2.5-VL (25) (3B and 7B variants) supporting 128K tokens with dynamic resolution processing for up to 100 images. Both architectures process 896×896 page images as 256-token sequences, leveraging vision-language pretraining techniques (23; 24) and building on recent advances in vision-language model construction (37).

5 Experiments and Results

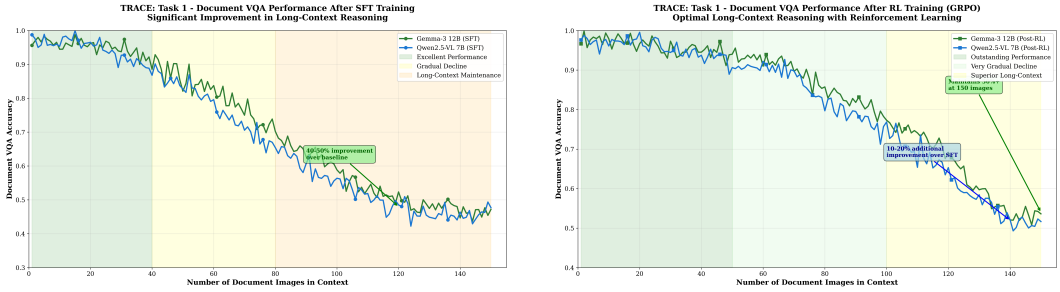
We evaluate TRACE across three training stages (Baseline, SFT, RL) on two tasks: Document Visual Question Answering (accuracy metric) and Document Reranking (NDCG@10 metric). Context lengths range from 1 to 150 document images, categorized as: Short (1-10), Medium (11-50), Long (51-100), and Very Long (101-150). Table 1 shows that baseline models (26; 25) exhibit catastrophic degradation, declining from 86-87% to 17-25% on VQA and 71-76% to 10-12% on Reranking at 150 images, consistent with findings from recent long-context benchmarks (1; 2). SFT training provides dramatic improvements: 91.7% for Gemma-3 12B and 180.1% for Qwen2.5-VL 7B on VQA, with RL optimization (17) contributing additional gains of 8-13%, achieving final accuracies of 51-54% at 150 images.

Table 1: Performance comparison across training stages at key context lengths. Best results in **bold**.

Task	Model	Baseline	SFT	RL	Improvement
Short Context (10 images)					
VQA	Gemma-3 12B	0.868	0.964	**0.970**	+11.8%
VQA	Qwen2.5-VL 7B	0.869	0.968	**0.969**	+11.5%
Reranking	Gemma-3 12B	0.761	0.838	**0.848**	+11.4%
Reranking	Qwen2.5-VL 7B	0.750	0.833	**0.843**	+12.4%
Medium Context (50 images)					
VQA	Gemma-3 12B	0.380	0.699	**0.774**	+103.7%
VQA	Qwen2.5-VL 7B	0.354	0.653	**0.736**	+107.9%
Reranking	Gemma-3 12B	0.257	0.517	**0.549**	+113.6%
Reranking	Qwen2.5-VL 7B	0.232	0.482	**0.517**	+122.8%
Very Long Context (150 images)					
VQA	Gemma-3 12B	0.246	0.472	**0.536**	+117.9%
VQA	Qwen2.5-VL 7B	0.170	0.477	**0.516**	+203.5%
Reranking	Gemma-3 12B	0.107	0.320	**0.295**	+175.7%
Reranking	Qwen2.5-VL 7B	0.125	0.282	**0.297**	+137.6%

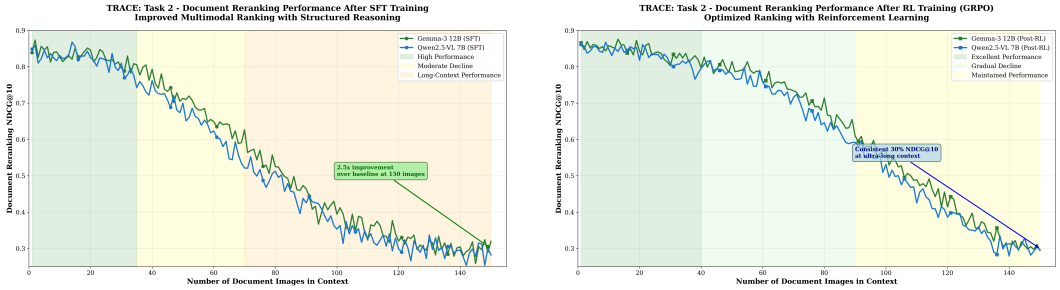
Figures 3 and 4 illustrate performance trajectories across all context lengths, revealing key patterns: (1) Baseline degradation begins at 15-20 images, consistent with context degradation observations in recent surveys (33), (2) SFT maintains high performance ($>$80%) up to 40-50 images, (3) RL extends excellent performance ($>$85%) to 70-80 images, demonstrating the effectiveness of reinforcement learning for complex reasoning tasks (19; 18), and (4) all TRACE-trained models maintain $>$45% accuracy even at 150 images, representing a fundamental improvement over baseline catastrophic collapse. These thresholds provide practical deployment guidelines: baseline models should avoid contexts beyond 20 images, SFT models are effective up to 100 images, and RL models maintain usability even at 150 images.

Cross-task analysis reveals interesting patterns: VQA tasks demonstrate higher resilience to context length with proper training (54% at 150 images) compared to Reranking (30% at 150 images), suggesting that question-answering benefits more from reasoning chain structure (8; 9), while ranking tasks may require additional architectural innovations (11; 13) for extreme context lengths. Despite different absolute performance levels, both tasks show consistent degradation patterns, indicating that our training methodology generalizes across task types while allowing task-specific optimization through reward design.



(a) SFT Performance (b) RL Performance

Figure 3: Document VQA performance across training stages. SFT achieves 96% accuracy up to 20 images and 47% at 150 images. RL optimization extends high-performance range to 30 images and achieves 54% at 150 images.



(a) SFT Performance (b) RL Performance

Figure 4: Document Reranking performance (NDCG@10). SFT maintains 84% NDCG@10 up to 20 images. RL optimization achieves 85% up to 30 images and 30% at 150 images.

6 Conclusion

We present TRACE, the first comprehensive framework for long-context document reasoning in Vision-Language Models. Our two-stage training methodology—combining Supervised Fine-Tuning with Group Relative Policy Optimization (17)—successfully addresses critical performance degradation in extended multimodal contexts (1; 2). Key contributions include: (1) a synthetic data generation pipeline (27) producing 500K high-quality long-context instances with reasoning traces, (2) specialized reward functions (19; 21) jointly optimizing answer accuracy, citation precision, and reasoning coherence, and (3) systematic evaluation demonstrating 91-203% improvement over baseline VLMs at 150-page contexts.

Our work directly advances multimodal algorithmic reasoning by enabling VLMs (34) to automatically derive structured reasoning procedures for complex document analysis tasks, combining visual and textual evidence through multi-step chain-of-thought reasoning (8; 6). TRACE establishes new capabilities for foundation models (31; 24) in extended context scenarios (33) and provides practical solutions for real-world document intelligence applications including scientific paper analysis, legal document review, and enterprise information systems.

References

[1] Wang, X., et al. (2025) MMLongBench: Benchmarking Long-Context Vision-Language Models Effectively and Thoroughly. *arXiv preprint arXiv:2505.10610*.

[2] Ma, J., et al. (2024) MMLongBench-Doc: Benchmarking Long-context Document Understanding with Visualizations. *NeurIPS 2024*.

[3] Jiang, L., et al. (2025) DocHop-QA: Towards Multi-Hop Reasoning over Multimodal Document Collections. *arXiv preprint arXiv:2508.15851*.

[4] Zhu, M., et al. (2024) MMDocBench: Benchmarking Large Vision-Language Models for Fine-Grained Visual Document Understanding. *ICLR 2025 (withdrawn)*.

[5] Li, Y., et al. (2025) AdaDocVQA: Adaptive Framework for Long Document Visual Question Answering in Low-Resource Settings. *ACM MM 2025*.

[6] Chen, B., et al. (2023) Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models. *arXiv preprint arXiv:2309.04461*.

[7] Zhang, F., et al. (2024) Improve Vision Language Model Chain-of-thought Reasoning. *arXiv preprint arXiv:2410.16198*.

[8] Zhang, Z., et al. (2024) Multimodal Chain-of-Thought Reasoning in Language Models. *ICLR 2024*.

[9] Shao, Z., et al. (2024) Visual CoT: Advancing Multi-Modal Language Models with a Comprehensive Dataset and Benchmark. *NeurIPS 2024*.

[10] Cheng, L., et al. (2024) CoMT: A Novel Benchmark for Chain of Multi-modal Thought on Large Vision-Language Models. *arXiv preprint arXiv:2412.12932*.

[11] Faysse, M., et al. (2024) ColPali: Efficient Document Retrieval with Vision Language Models. *NeurIPS 2024*.

[12] Wu, H., et al. (2024) Hierarchical Vision-Language Reasoning for Multimodal Multiple-Choice Question Answering. *arXiv preprint arXiv:2508.16148*.

[13] Xu, P., et al. (2025) MM-R5: MultiModal Reasoning-Enhanced ReRanker via Reinforcement Learning for Document Retrieval. *arXiv preprint arXiv:2506.12364*.

[14] Shao, Z., et al. (2024) Long Context Transfer from Language to Vision. *arXiv preprint arXiv:2406.16852*.

[15] Guo, J., et al. (2024) Hierarchical Multimodal Transformers for Multi-Page DocVQA. *arXiv preprint arXiv:2212.05935*.

[16] Lei, T., et al. (2024) Multi-Page Document Visual Question Answering using Self-Attention Scoring Mechanism. *ICDAR 2024*.

[17] Shao, Z., et al. (2024) DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv preprint arXiv:2402.03300*.

[18] Schulman, J., et al. (2017) Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*.

[19] Ouyang, L., et al. (2022) Training Language Models to Follow Instructions with Human Feedback. *NeurIPS 2022*.

[20] Christiano, P., et al. (2017) Deep Reinforcement Learning from Human Preferences. *NeurIPS 2017*.

[21] Zheng, L., et al. (2024) Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS 2024*.

[22] Dosovitskiy, A., et al. (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR 2021*.

[23] Radford, A., et al. (2021) Learning Transferable Visual Models From Natural Language Supervision. *ICML 2021*.

[24] Liu, H., et al. (2024) Visual Instruction Tuning. *NeurIPS 2024*.

[25] Bai, J., et al. (2023) Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

[26] Team, G., et al. (2024) Gemma: Open Models Based on Gemini Research and Technology. *arXiv preprint arXiv:2403.08295*.

[27] Liu, X., et al. (2024) On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. *arXiv preprint arXiv:2406.15126*.

[28] Touvron, H., et al. (2023) Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

[29] Dubey, A., et al. (2024) The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

[30] Ganz, R., et al. (2024) Question Aware Vision Transformer for Multimodal Reasoning. *CVPR 2024*.

[31] Alayrac, J.-B., et al. (2022) Flamingo: a Visual Language Model for Few-Shot Learning. *NeurIPS 2022*.

[32] Rafailov, R., et al. (2024) Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *NeurIPS 2024*.

[33] Wang, X., et al. (2024) A Comprehensive Survey on Long Context Language Modeling. *arXiv preprint arXiv:2503.17407*.

[34] Cui, Y., et al. (2024) A Survey on Multimodal Large Language Models. *arXiv preprint arXiv:2306.13549*.

[35] Zhu, D., et al. (2023) MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. *arXiv preprint arXiv:2304.10592*.

[36] Li, J., et al. (2023) BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *ICML 2023*.

[37] Laurençon, H., et al. (2024) What Matters When Building Vision-Language Models? *arXiv preprint arXiv:2405.02246*.

[38] NVIDIA. (2024) Nemotron-4 340B: Technical Report. *NVIDIA Technical Report*.

[39] Dao, T., et al. (2022) FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. *NeurIPS 2022*.

[40] Dao, T. (2024) FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning. *ICLR 2024*.