STA-CoT: Structured Target-Centric Agentic Chain-of-Thought for Consistent Multi-Image Geological Reasoning

Anonymous ACL submission

Abstract

Reliable multi-image geological reasoning is essential for automating expert tasks in remote-sensing mineral exploration, yet remains challenging for multimodal large language models (MLLMs) due to the need for locating target areas, accurate cross-image referencing, and consistency over long reasoning chains. We propose STA-CoT, a Structured Target-centric Agentic Chain-of-Thought framework that orchestrates planning, execution, and verification agents to decompose, ground, and iteratively refine reasoning steps over geological and hyperspectral image sets. By aligning each reasoning step to specific image target areas and enforcing consistency through agentic verification and majority voting, STA-CoT robustly mitigates tool errors, long-chain inconsistencies, and error propagation. We rigorously evaluate STA-CoT on MineBench, a dedicated benchmark for multi-image mineral exploration, demonstrating substantial improvements over existing multimodal chain-of-thought and agentic baselines. Our results establish STA-CoT as a reliable and robust solution for consistent multi-image geological reasoning, advancing automated scientific discovery in mineral exploration.

1 Introduction

Multi-image geological reasoning is the task of synthesizing and interpreting geological evidence from multiple remote-sensing images of a region to produce a cross-image rationale over targeted areas in the region (Alzubaidi et al., 2021). Among its most impactful applications is mineral exploration (Yousefi et al., 2019), where experts integrate geological and hyperspectral imagery to identify and assess economically valuable mineral deposits. The ability to automate mineral exploration holds tremendous significance, as it accelerates resource discovery essential for global technological infrastructure and sustainability (Sabins, 1999). A domain expert may spend hours manually correlating structured faults, alteration zones, and spectral signatures across several images to narrow down large survey regions into a small set of promising candidates – demonstrating how automation can drastically reduce the number of regions requiring costly field investigation (Shirmard et al., 2022).

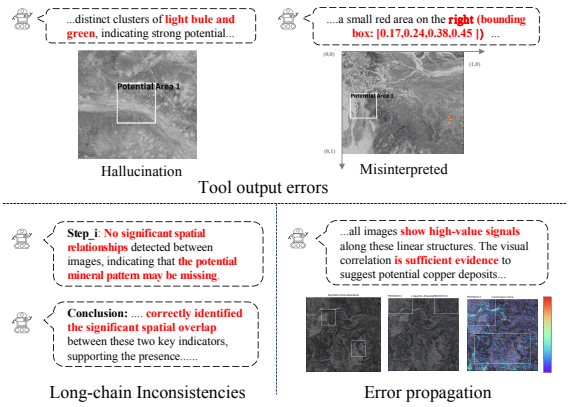


Figure 1: Three types of errors in visual-based augmentation approaches in multi-image geological reasoning.

The performance of multi-image geological reasoning fundamentally relies on a model's ability to reason across both spatial and spectral modalities-integrating diverse geological cues distributed among multiple images over one region into a coherent, evidence-based conclusion. This motivates the adoption of multimodal large language models (MLLMs), which unify visual and textual inputs, as well as the multimodal chain-of-thought (MCoT) framework that decomposes complex visual reasoning into interpretable, step-by-step rationales (Wang et al., 2025a). Recent advances have further strengthened these reasoning frameworks by introducing prompt-based MCoT (Fei et al., 2024; Zhang et al., 2024), structured planning (Wei et al., 2024; Wang et al., 2025b), tool-augmented decision-making (Tang et al.; Wu et al., 2024), and mechanisms for consistency (Zhou et al., 2024) and verification (Yan et al., 2025). Such techniques en-

able MLLMs to not only elaborate their reasoning processes but also actively validate and refine their intermediate outputs, leading to accurate answers.

However, these techniques cannot be directly applied to multi-image geological reasoning, as, besides proper geological knowledge, the task uniquely demands (i) accurate interesting target location within each image, (ii) maintenance of a long and complex reasoning chain, and (iii) precise cross-image referencing that exceeds the typical capabilities of existing approaches. In our pilot experiments, when we adopt MCoT methods with tool-augmentation and vanilla consistency & verification for this task, we observed three primary categories of failure as in Figure 1: tool output errors, long-chain inconsistencies, and error propagation. These errors arise precisely because this task amplifies the risk of faulty target location by tools due to a lack of domain knowledge, increases the chance for intermediate mistakes to accumulate and propagate along extended reasoning chains, and makes it difficult to maintain consistent references and logic across multiple images – ultimately undermining the reliability of the conclusions.

Motivated by these challenges, we seek to enable reliable and domain-adapted multi-image reasoning for geological exploration. To this end, we present Structured Target-centric Agentic Chain-of-Thought, STA-CoT, a novel framework designed to deliver consistent and accurate multi-image geological reasoning. It orchestrates specialized agentic modules within a structured chain-of-thought paradigm, explicitly aligning reasoning steps with domain knowledge, visual targets, and cross-image dependencies. By structuring the reasoning process around targeted regions and incorporating agent-based planning, execution, and verification, STA-CoT mitigates the unique complexities of geological multi-image tasks. Specifically, it coordinates three closely integrated agents: a Planner that retrieves relevant geological knowledge and decomposes the complex multimodal task into a sequence of manageable sub-tasks; an Executor that performs these sub-tasks by invoking specialized tools and integrates outputs via a rule-based controller to ensure accurate visual alignment; and a Verifier that continuously monitors intermediate reasoning steps, triggering a stepwise refinement process to direct the Executor to revise faulty outputs, thereby preventing error propagation and maintaining causal stability throughout the chain.

As such, STA-CoT grounds each reasoning step to specific image regions and systematically tracks cross-image relationships through structured agentic planning. Consistency is maintained by continually verifying intermediate outputs, while a majority voting mechanism further reduces long-chain inconsistencies, yielding stable and reliable conclusions in complex multi-image geological reasoning. To rigorously evaluate our approach, we leverage MineBench, a dedicated benchmark designed for multi-image mineral exploration tasks, which integrates both geological and hyperspectral remote-sensing data to emulate real-world exploration scenarios (Yu et al., 2024). Empirical results demonstrate that STA-CoT achieves state-of-the-art performance, substantially outperforming prior MCoT, multi-agent, and tool-augmented MLLM baselines. These affirm the framework's effectiveness in delivering reliable and consistent geological reasoning across complex multi-image contexts.

2 Related Work

Multimodal Chain-of-Thought. Multimodal Chain-of-Thought (MCoT) (Zhang et al., 2023) reasoning frameworks have rapidly advanced from early prompt-based works that generate stepwise textual rationales grounded in visual content (Fei et al., 2024; Zhang et al., 2024) to sophisticated paradigms like retrieved-augmented, structured and tool-augmented multimodal reasoning. Retrieval-Augmented Generation (RAG) injects external or domain-specific knowledge to inform inferences in complex scenarios (Dong et al., 2024; Pan et al., 2024). And, structured and planning-based reasoning approaches (Gao et al., 2024; Hu et al., 2024), such as Graph-of-Thought (GoT) (Besta et al., 2024) and Compositional CoT (CCoT) (Mitra et al., 2024), explicitly model the dependencies and relationships between reasoning steps, often leveraging graph-based or compositional representations to capture intricate cross-modal interactions. Meanwhile, agentic tool-augmented and Chain-of-Action (CoA) methods, including KAM-CoT (Mondal et al., 2024), MM-Verify (Sun et al., 2025), and Det-CoT(Wu et al., 2024), extend MCoT by integrating external tools (e.g., visual annotators, object detection, and recognition) to enable interactive perception and iterative refinement of visual evidence. Despite these advances, most existing paradigms are evaluated on single-image or short-context tasks, lack robust mechanisms for explicitly tracking cross-image

references. In contrast, STA-CoT is designed to tackle the unique challenges of multi-image geological reasoning, where retrieved-augmented domain-contextual planning, accurate alignment and referencing of targeted regions across images, and structured management of long and error-prone reasoning chains are indispensable for consistent decision making.

Consistency and Verification. Maintaining logical consistency is particularly challenging in safety-critical or scientific domains like mineral exploration, where errors at intermediate reasoning steps can undermine overall trustworthiness (Havrilla and Iyer, 2024). Existing approaches to consistency in reasoning generally operate at the chain or output level, employing strategies such as majority vote (Tan et al., 2024; Zelikman et al., 2022) and post-hoc verification (Sun et al., 2025; Yan et al., 2025). While these methods can improve the plausibility of final predictions by selecting among alternative reasoning chains, they offer limited mechanisms for diagnosing and repairing errors within the reasoning process itself (Zhou et al., 2024). Such coarse-grained consistency enforcement is often insufficient in domains where interpretability and reliability at every reasoning step are required. In contrast, our STA-CoT framework addresses this critical gap by introducing an agent-driven, iterative refinement mechanism that enables stepwise error detection and targeted repair, thus supporting robust and trustworthy multimodal reasoning.

Mineral Exploration. Mineral exploration exemplifies multi-image geological reasoning, requiring the integration of geological and hyperspectral evidence from multiple remote-sensing images to identify mineralization patterns and predict deposit locations (Sabins, 1999; Zuo et al., 2021). Unlike standard multimodal tasks, models must accurately detect and correlate key geological features across images, mirroring expert geoscientists' approach of synthesizing spatial and spectral cues for reliable decisions. Although agentic MLLMs have advanced scientific reasoning in fields such as mathematics (Deng et al., 2024), medicine (Kim et al., 2024; Li et al., 2024), and geoscience (Liu et al., 2024; Xu et al., 2024), few frameworks (Yu et al., 2024) address the specific challenge of cross-image reasoning essential to mineral exploration. This motivates our STA-CoT framework, which explicitly structures agentic reasoning around multi-image evidence to achieve consistent and robust mineral prospectivity assessments. In contrast to MineAgent, which primarily established modular pipelines and benchmarks, this work presents a structured agentic chain-of-thought framework that directly tackles cross-image consistency and targeted geological reasoning.

3 Methodology

To enable automated mineral exploration with remote-sensing imagery, we formally define the reasoning objective as follows.

Task Formulation. Given a mineral-exploring query Q (e.g., *'Given the following remote-sensing images, can deposits be detected? Answer: A. Yes B. No.'*) and a set of remote-sensing images $\mathcal{I}$ representing a targeted region, the objective is to determine the presence or absence of deposits:

$$A_{\text{final}} \sim M(\mathcal{I}, Q; \theta_M) \tag{1}$$

where M denotes a multimodal large language model parameterized by θ_M, and $A_{\text{final}} \in \{A, B\}$ is the final predicted answer.

To address the complexity of multi-image geological reasoning – requiring integration of spatial and spectral evidence, cross-image referencing, and logical consistency – we adopt an agentic reasoning framework, as inspired by prior works (Yan et al., 2025; Li et al., 2024; Kim et al., 2024; Sun et al., 2025). Within this framework, the model not only produces a final answer but also outputs an interpretable reasoning chain:

$$R = (r^{(1)}, \ldots, r^{(N)}) \tag{2}$$

where each step $r^{(i)}$ consists of a structured subtask and its corresponding multimodal observation over N steps. Furthermore, STA-CoT explicitly orchestrates structured domain-informed planning (§3.1), target-centric visual execution (§3.2), and agent-driven consistency verification (§3.3) to handle geological reasoning tasks as below. Please see Figure 2 for an illustration.

3.1 Structured Domain-informed Planning

Effective multi-image geological reasoning demands not only general visual understanding but also the ability to decompose complex queries into domain-specific, actionable steps. To this end, STA-CoT incorporates a domain-informed planner (M_p), which systematically translates each mineral

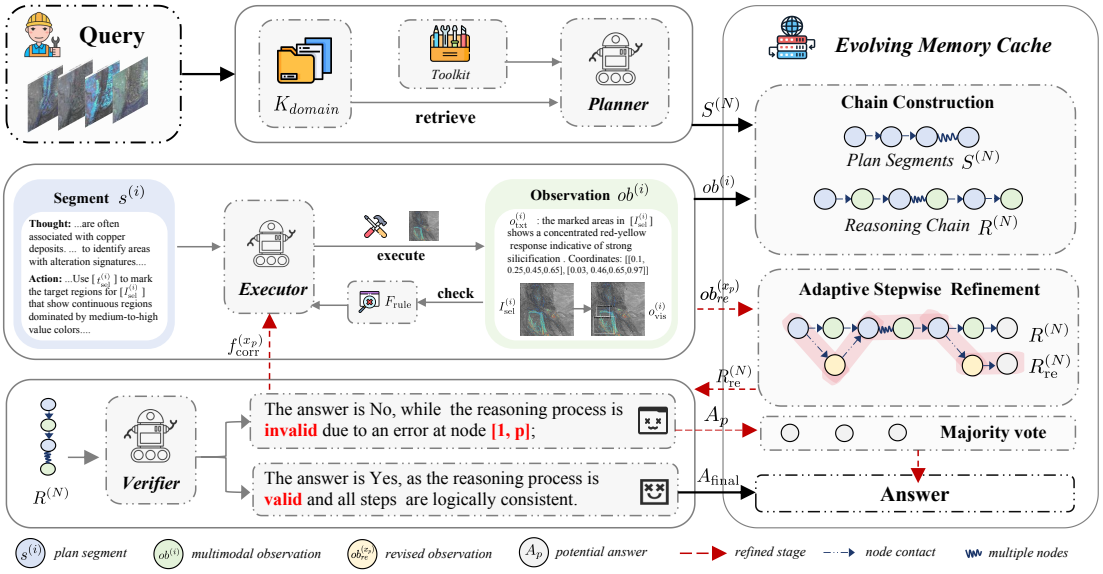


Figure 2: Our Structured Target-centric Agentic Chain-of-Thought (STA-CoT) framework for multi-image geological reasoning, which consists of a Planner (M_p), an Executor (M_e), and a Verifier (M_v).

exploration query Q into a structured sequence of plan segments $S = (s^{(1)}, \ldots, s^{(N)})$.

This planner operates by leveraging a domain knowledge base (K_{domain}), encapsulating key geological concepts, and a curated visual toolkit (T_{set}) tailored to mineral exploration. The toolkit provides specialized operations, such as (i) targeting alteration zones (*box maker color mode*), (ii) identifying salient geological structures (*box maker feature region*), (iii) exploring spatial relationships across images (*spatial relationship explorer*), and (iv) integrating evidence for decision-making, as detailed in Appendix C.4.

By orchestrating these components, M_p generates visually grounded, executable plans that break down the task into sub-steps, each aligned with relevant image area and geological priors:

$$S \sim M_p(Q, \mathcal{I}, K_{\text{domain}}, T_{\text{set}}; \theta_p) \tag{3}$$

where θ_p are the model parameters. This structured, domain-informed planning forms the foundation for accurate multi-image geological reasoning.

3.2 Target-centric Visual-grounded Execution

To faithfully realize each step of the structured reasoning plan, STA-CoT employs a model-based executor (M_e) that transforms abstract plans into concrete multimodal actions. Unlike generic execution, our executor explicitly aligns every sub-task with precise target areas across the input images, ensuring that each operation is spatially grounded and contextually relevant.

For each plan segment $s^{(i)}$, the executor identifies the corresponding target area within the selected imagery $\mathcal{I}_{\text{sel}}^{(i)}$, then invokes the most appropriate visual tool $t_{\text{sel}}^{(i)}$ as specified by the planner. This process yields a rich observation $ob^{(i)} = (o_{\text{txt}}^{(i)}, o_{\text{vis}}^{(i)})$, consisting of both a textual rationale ($o_{\text{txt}}^{(i)}$) and a visual outcome ($o_{\text{vis}}^{(i)}$), such as an annotated or segmented image.

Crucially, our execution is target-centric – all actions and tool invocations are explicitly conditioned on the spatial context of geological interest, which is vital for accurate cross-image reasoning. Moreover, the process is visual-grounded – ensuring that each step produces verifiable visual evidence.

To safeguard the fidelity of each execution, M_e incorporates an internal rule-based controller F_{rule}, which checks the alignment between described regions and visual outputs, automatically flagging and correcting inconsistencies (see Appendix C.3). This mechanism, together with optional agentic feedback $[f_{\text{ext}}^{(k)}]$, promotes error-resistant execution:

$$ob^{(i)} = (o_{\text{txt}}^{(i)}, o_{\text{vis}}^{(i)}) \sim M_e(s^{(i)}, [f_{\text{ext}}^{(k)}]; \theta_e) \tag{4}$$

where θ_e denotes executor parameters.

By explicitly grounding each reasoning step in both targeted spatial context and visual evidence, the STA-CoT executor ensures that multi-image geological reasoning is interpretable, traceable, and resilient to cascading errors.

3.3 Agentic Consistency-driven Verification

Long-chain multi-image geological reasoning is uniquely prone to error propagation and logical inconsistency, as minor mistakes in intermediate steps can accumulate and distort the final conclusion. Traditional output-level consistency mechanisms are insufficient for safety-critical scientific domains, where stepwise interpretability and reliability are paramount. Therefore, STA-CoT introduces an agentic verifier to proactively monitor, validate, and iteratively repair the reasoning process, ensuring robust and trustworthy outcomes.

Verifier Basic. The verifier (M_v) serves as a critical agent in STA-CoT, tasked with examining the entire reasoning chain $R = (r^{(1)}, \ldots, r^{(N)})$, where each node $r^{(i)} = (s^{(i)}, ob^{(i)})$ links a planned segment and its multimodal observation. M_v conducts fine-grained checks at both node and chain levels—validating logical soundness, factual correctness, and semantic alignment between textual and visual components (e.g., between $o^{(i)}_{\text{txt}}$ and $o^{(i)}_{\text{vis}}$). On detecting flaws, the verifier generates targeted corrective feedback F_{corr} for self-repair. The formal output of the verifier is:

$$(\text{isValid}, A_{\text{p}}, F_{\text{corr}}) \sim M_v(R, Q; \theta_v) \qquad (5)$$

where isValid flags chain validity, A_{p} is the candidate answer, and F_{corr} lists corrective guidance.

Progressive Reasoning Chain Construction. STA-CoT utilizes a memory cache M_{cache} as a shared workspace initialized with the query Q and input images $\mathcal{I}$, continuously enriched with intermediate results. The planner M_p generates a structured sequence of plan segments $S^{(N)}$, while the executor M_e performs each $s^{(i)}$, producing multimodal observations $ob^{(i)}$. Each resulting reasoning node $r^{(i)} = (s^{(i)}, ob^{(i)})$ is appended to the chain:

$$R^{(i)} \leftarrow R^{(i-1)} \oplus r^{(i)} \qquad (6)$$

This progressive execution and cache updating ensure every step leverages the latest context, facilitating coherent and context-aware reasoning.

Adaptive Stepwise Chain Refinement. After constructing $R^{(N)}$, the verifier M_v checks validity as per Eq. 5. If isValid $=$ true, A_p is accepted as the final answer A_{final}. If isValid $=$ false, M_v identifies m erroneous nodes in $R^{(N)}$, indexed by $\{x_1, \ldots, x_m\}$, and issues targeted revision guidance $f^{(x_p)}_{\text{corr}}$ for each. These nodes are then re-executed by the executor M_e using the feedback, generating revised observations $ob^{(x_p)}_{\text{re}}$, and updating the chain as follows:

$$r^{(x_p)}_{\text{re}} \leftarrow (s^{(x_p)}, ob^{(x_p)}_{\text{re}}) \qquad (7)$$

The refined chain $R^{(N)}_{\text{re}}$ is re-validated by M_v, and this loop continues until validation succeeds or a maximum number of iterations is reached.

Fallback Global Majority Vote. If the chain fails to pass verification after K rounds, STA-CoT employs a majority voting mechanism. For each round $r \in [1, K]$, a candidate answer $A^{(r)}_p$ is extracted from the refined chain $R^{(N),r}_{\text{re}}$. The final answer is then determined as:

$$A_{\text{final}} \leftarrow \text{majority-vote}\left(A^{(1)}_p, \ldots, A^{(K)}_p\right) \qquad (8)$$

This approach ensures robust decision-making by aggregating the most consistent answer across multiple correction attempts.

By integrating fine-grained, agentic verification with iterative repair and robust fallback voting, STA-CoT delivers consistent, interpretable, and scientifically trustworthy multi-image geological reasoning, as detailed in Algorithm A.

4 Experiment

Implementation Details. We evaluate STA-CoT[1] on the MineBench dataset (Yu et al., 2024), which is a recently proposed geological reasoning task focused on multi-image mineral deposit identification. Following (Yu et al., 2024), we report three evaluation metrics: F1 score for the positive class (Pos.F1), macro-averaged F1 (Avg.F1), and Matthews Correlation Coefficient (MCC) (Chicco and Jurman, 2020). STA-CoT consists of three core roles: the planner, executor, and verifier, where the planner is instantiated with the Gemini-2.0 (Team et al., 2024), as high-quality planning is essential for guiding the execution and verification of complex multi-image tasks. The executor and verifier roles are realized using a diverse set of MLLMs, including proprietary models such as GPT-4o (OpenAI, 2024) and Gemini-2.0, and the open-source Qwen-2.5-7B (Bai et al., 2025). The detailed experiment setting and prompts for three modules are provided in Appendix C.1 and C.4.

[1] Our anonymous code is available at https://anonymous.4open.science/r/STA-CoT/.

Model	Method	Metrics		
		Pos.F1	Avg.F1	MCC
GPT-4o	MCoT	34.81	57.19	26.07
	RAG	35.92	58.32	27.46
	Tool-augmented	26.01	33.38	17.66
	MineAgent	61.20	77.19	56.30
	STA-CoT	**63.10**	**78.62**	**58.03**
Gemini-2.0	MCoT	37.50	66.10	40.45
	RAG	41.12	67.75	39.51
	Tool-augmented	37.50	55.16	33.97
	MineAgent	60.18	78.06	59.62
	STA-CoT	**66.67**	**80.75**	**62.12**

Table 1: Performance comparison of different MCoT methods on the MineBench benchmark.

Method	Pos.F1	Avg.F1	MCC
w/o Planner	52.05	72.79	45.59
w/o Executor	10.39	52.20	22.04
w/o Verifier	37.50	55.16	33.97
STA-CoT	**66.67**	**80.75**	**62.12**

Table 2: Ablation Study using Gemini-2.0.

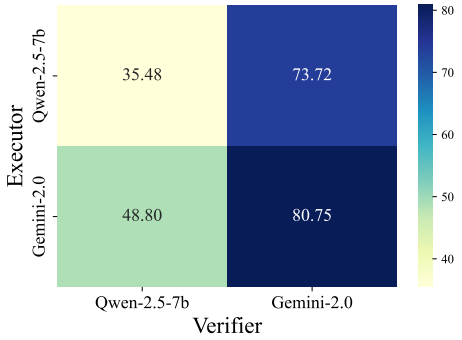


Figure 3: Performance for different back-end model combinations of Executor and Verifier (Avg.F1).

4.1 Main Result

In Table 1, we evaluate STA-CoT against four representative baselines as identified in Section 2: standard MCoT, RAG, Tool-augmented, and MineAgent (structured and planning-based reasoning) methods. The tool-augmented baseline is implemented using our target-centric tools without additional verification or refinement steps. RAG yields only marginal improvement over standard MCoT, while tool-augmented baseline degrades performance due to the injection of noisy or unvalidated relational information in the absence of verification. MineAgent achieves substantial gains over unimodal and less structured baselines. STA-CoT achieves the highest scores across all evaluation metrics, surpassing MineAgent by +1.4% Avg.F1 on GPT-4o and +2.6% Avg.F1 on Gemini-2.0. These results underscore the importance of structured planning and consistent reasoning.

4.2 Necessity of Each Module

STA-CoT explicitly modularizes Planner, Executor, and Verifier. To validate the complementary roles of these modules, we first conduct ablation studies to measure the contributions of each component. This analysis identifies how each component addresses key challenges such as evidence integration, error suppression, and logical consistency.

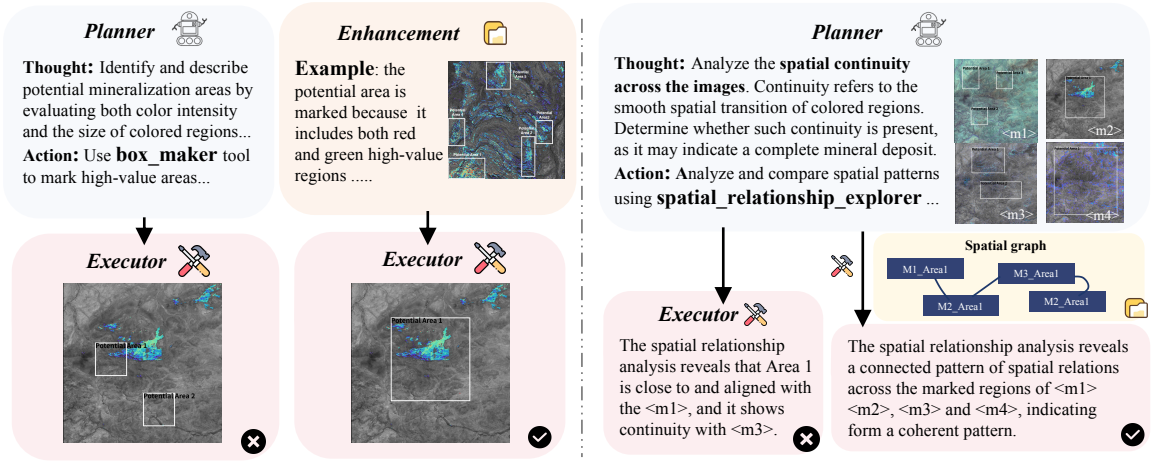
As shown in Table 2, removing domain knowledge by disabling K_{domain} in the Planner leads to a noticeable, but relatively moderate, reduction in performance. This suggests that explicit domain adaptation is necessary for well-structured and contextually relevant plans, yet the Planner's impact is less critical as MLLMs have been equipped with fair geological knowledge. In contrast, ablating the Executor, thus eliminating multimodal evidence extraction, results in a substantial drop, highlighting the Executor's essential role in providing visual grounding. Similarly, disabling the Verifier causes a marked Avg.F1 decline, underscoring its pivotal function in maintaining logical reliability and suppressing error propagation.

These results confirm that while all three modules contribute to overall reasoning performance, the Executor and Verifier are particularly indispensable and complementary.

4.3 Efficient Executor-Verifier Tradeoff

As it's verified that our Executor and Verifier play dominant roles in the above ablation study, we further investigate how their assignment and capacity trade-offs impact overall performance and system efficiency. The Executor is invoked at every reasoning step for stepwise multimodal evidence extraction, making it a candidate for lightweight model deployment to reduce computational cost. In contrast, the Verifier operates only at the chain level, enforcing global consistency and error correction, and thus can be allocated higher-capacity models without incurring significant resource overhead.

Empirical results in Figure 3 validate this design: increasing either module's capacity improves performance, but upgrading the Verifier yields a more pronounced gain than upgrading the Executor. Notably, pairing a strong Verifier with a weaker Executor achieves significantly higher Avg.F1 than



(a) w/o vs. w/ Target-centric Visual Guidance. **(b)** w/o vs. w/ Graph-based Spatial Reasoning.

Figure 4: Comparison of Qwen-2.5-7b performance with and without visual-grounded enhancements within Executor: (a) visual guidance in the box maker tool; (b) graph-based guidance in the spatial relationship explorer.

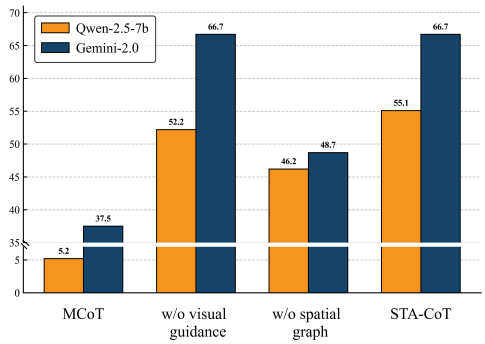


Figure 5: Impact of visual-grounded enhancements on Pos.F1 scores for Qwen-2.5-7b and Gemini-2.0.

the reverse, illustrating that robust chain-level verification can effectively compensate for upstream limitations in evidence extraction. These findings provide a practical insight into STA-CoT: in real-world deployments, prioritizing resources for the Verifier while using efficient models for execution can deliver strong reasoning performance with minimal computational cost.

4.4 Executor: Enhancing Visual-grounding

Motivated by the core challenges of multi-image geological reasoning, particularly the need for accurate target identification and precise cross-image referencing, we evaluate the effectiveness of visual-grounded augmentation within the Executor.

Visual Guidance Enhances Target Accuracy. To address the challenge of accurate localization within each geological image, we integrate visual guidance into the box maker tool, providing curated annotation examples and domain-specific prompts. As shown in Figure 4(a), removing visual guidance leads the Executor to produce noisy or erroneous annotations, which yields a significant improvement in Pos.F1 scores (Figure 5), confirming its critical role in accurate geological information.

Graph-based Reasoning Improves Cross-image Consistency. To effectively maintain precise cross-image referencing, another critical challenge, we incorporate symbolic graph-based relational reasoning into the spatial relationship explorer tool (see Appendix C.2). The graph-based augmentation explicitly encodes relational information across multiple images, facilitating the continuity and consistency that MLLMs alone struggle to achieve (Figure 4(b)). Removing this mechanism results in a sharp performance decline (Figure 5), emphasizing that graph-based reasoning is essential for capturing complex spatial dependencies.

Essential for Weaker MLLMs. These visual-grounded mechanisms are particularly beneficial for smaller models such as Qwen-2.5-7B, substantially improving their baseline accuracy. Larger models, such as Gemini-2.0, also benefit from structured visual and relational guidance, further confirming the general applicability and necessity of these enhancements across models.

4.5 Verifier: Enhancing Chain Consistency through Stepwise and Global Correction

Given the complexity of long-chain geological reasoning, we evaluate two key consistency mechanisms integrated within Verifier: stepwise refinement and global majority voting (Fig. 6). Both

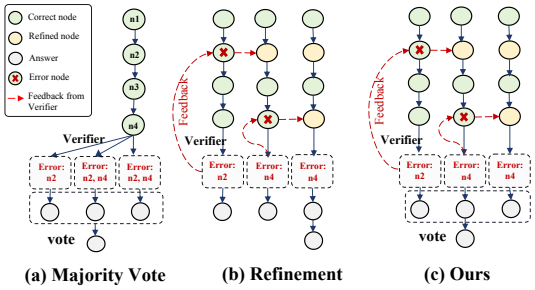


Figure 6: Comparison of consistency mechanisms. (a) Majority vote, (b) Stepwise refinement, (c) STA-CoT.

Method	Pos.F1	Avg.F1	MCC
w/o refinement	51.30	68.81	49.31
w/o majority vote	35.05	64.74	36.73
STA-CoT (full)	**66.67**	**80.75**	**62.12**

Table 3: Ablation results for stepwise refinement and global majority vote mechanisms on Gemini-2.0.

mechanisms individually and in combination contribute to improved reasoning reliability and accuracy, as reported in Table 3.

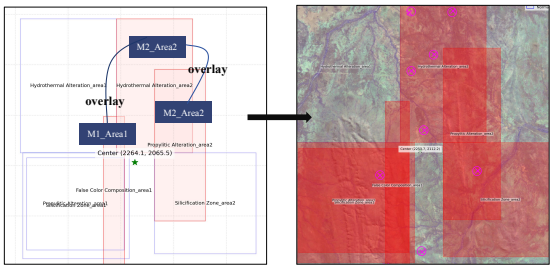
Global Majority Voting Stabilizes Outputs. Employing global majority voting without refinement partially stabilizes outputs by aggregating multiple independent answers from a reasoning path. However, early errors often persist and propagate throughout the chain, resulting in only moderate performance. This indicates that majority voting alone is insufficient to resolve correlated or persistent errors.

Stepwise Refinement Repairs Local Errors. Applying stepwise refinement without global voting effectively addresses some local errors within individual reasoning chains. However, this approach struggles with global inconsistencies across chains, which significantly reduces overall performance. Such unresolved inconsistencies highlight the necessity of global consensus mechanisms to achieve reliable reasoning.

STA-CoT achieves the best overall performance, confirming that the combination of adaptive local repair and global consensus enhances stability and interoperability of the geological reasoning.

4.6 Cross-images Deposit Localization

While STA-CoT primarily addresses classification-style reasoning, practical mineral exploration requires accurate localization of deposits that typically exhibit spatial continuity and extend across



(a) Spatial graph **(b)** Deposit localization

Figure 7: Deposit localization and area estimation. (a) Spatial graph illustrating candidate regions (red) and detected spatial relationships (blue). (b) Predicted high-potential regions are shown in red, while purple markers indicate ground-truth deposit locations.

multiple images. This necessitates cross-image reasoning to integrate fragmented evidence into spatially coherent mineralization zones.

To meet this need, we augment STA-CoT with a post-hoc localization module (Appendix C.4) that leverages the spatial graph constructed during execution, where nodes denote candidate regions and edges capture spatial relationships (e.g., overlap, adjacency), providing a structured basis for cross-image evidence aggregation. Then, the localization module refines candidate regions by iteratively merging spatially connected areas, emphasizing consistency across images and filtering out isolated or low-confidence predictions.

As shown in Figure 7, this approach enables spatially precise predictions closely aligned with ground truth. Empirically, it achieves a recall of 76.71% while reducing the explored area by 64.76%, demonstrating that structured cross-image reasoning is essential for accurate and efficient deposit localization in real-world geological tasks.

5 Conclusion

We proposed STA-CoT, a structured agentic reasoning framework for consistent multi-image geological reasoning in mineral exploration. STA-CoT integrates domain-informed planning, target-centric execution, and iterative verification to address cross-image dependencies and long-chain inconsistencies. Evaluated on the MineBench benchmark, STA-CoT outperforms prior methods in both accuracy and consistency, particularly excelling in visual-grounded execution and stepwise error correction. Our results demonstrate the framework's robustness, efficiency, and practical value for automating expert-level geological reasoning using remote-sensing data.

Limitations

One limitation of our current framework is the increased computational overhead introduced by multi-step execution and iterative refinement, which may result in higher inference latency and resource usage. Future work will explore optimization strategies to enhance computational efficiency while preserving reasoning robustness.

Additionally, the toolkit design is tailored to domain-specific characteristics, and the scalability of STA-CoT to larger-scale or cross-domain multi-image reasoning tasks remains to be systematically validated. We plan to systematically evaluate and adapt our approach for broader application scenarios in future research.

References

Fatimah Alzubaidi, Peyman Mostaghimi, Pawel Swietojanski, Stuart R Clark, and Ryan T Armstrong. 2021. Automated lithology classification from drill core images using convolutional neural networks. *Journal of Petroleum Science and Engineering*, 197:107933.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and 1 others. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17682–17690.

Davide Chicco and Giuseppe Jurman. 2020. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.

Linger Deng, Yuliang Liu, Bohan Li, Dongliang Luo, Liang Wu, Chengquan Zhang, Pengyuan Lyu, Ziyang Zhang, Gang Zhang, Errui Ding, and 1 others. 2024. R-cot: Reverse chain-of-thought problem generation for geometric reasoning in large multimodal models. *arXiv preprint arXiv:2410.17885*.

Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2024. Progressive multimodal reasoning via active retrieval. *arXiv preprint arXiv:2412.14835*.

Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, and 1 others. 2024. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105.

Alex Havrilla and Maia Iyer. 2024. Understanding the effect of noise in llm training data with algorithmic chains of thought. *arXiv preprint arXiv:2402.04004*.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, and 1 others. 2024. Mmedagent: Learning to use medical tools with multi-modal agent. *arXiv preprint arXiv:2407.02483*.

Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhenwei Shi. 2024. Change-agent: Towards interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431.

Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. 2024. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18798–18806.

OpenAI. 2024. Chatgpt-4o. Available at https://openai.com.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. 2024. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*.

Floyd F Sabins. 1999. Remote sensing for mineral exploration. *Ore geology reviews*, 14(3-4):157–183.

Hojat Shirmard, Ehsan Farahbakhsh, R Dietmar Müller, and Rohitash Chandra. 2022. A review of machine learning in processing remote sensing data for mineral exploration. *Remote Sensing of Environment*, 268:112750.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*.

Cheng Tan, Jingxuan Wei, Zhangyang Gao, Linzhuang Sun, Siyuan Li, Ruifeng Guo, Bihui Yu, and Stan Z Li. 2024. Boosting the power of small multimodal reasoning models to match larger models with self-consistency training. In *European Conference on Computer Vision*, pages 305–322. Springer.

J Tang, G Zheng, J Yu, and S Yang. Cotdet: Affordance knowledge prompting for task driven object detection. arxiv 2023. *arXiv preprint arXiv:2309.01093*.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025a. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Ziyue Wang, Junde Wu, Chang Han Low, and Yueming Jin. 2025b. Medagent-pro: Towards multi-modal evidence-based medical diagnosis via reasoning agentic workflow. *arXiv preprint arXiv:2503.18968*.

Lai Wei, Wenkai Wang, Xiaoyu Shen, Yu Xie, Zhihao Fan, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2024. Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration. *arXiv preprint arXiv:2410.04521*.

Yixuan Wu, Yizhou Wang, Shixiang Tang, Wenhao Wu, Tong He, Wanli Ouyang, Philip Torr, and Jian Wu. 2024. Dettoolchain: A new prompting paradigm to unleash detection ability of mllm. In *European Conference on Computer Vision*, pages 164–182. Springer.

Wenjia Xu, Zijian Yu, Yixu Wang, Jiuniu Wang, and Mugen Peng. 2024. Rs-agent: Automating remote sensing tasks through intelligent agents. *arXiv preprint arXiv:2406.07089*.

Yibo Yan, Shen Wang, Jiahao Huo, Philip S Yu, Xuming Hu, and Qingsong Wen. 2025. Mathagent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. *arXiv preprint arXiv:2503.18132*.

Mahyar Yousefi, Oliver P Kreuzer, Vesa Nykänen, and Jon MA Hronsky. 2019. Exploration information systems–a proposal for the future use of gis in mineral exploration targeting. *Ore Geology Reviews*, 111:103005.

Beibei Yu, Tao Shen, Hongbin Na, Ling Chen, and Denqi Li. 2024. Mineagent: Towards remote-sensing mineral exploration with multimodal large language models. *arXiv preprint arXiv:2412.17339*.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *arXiv preprint arXiv:2401.02582*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. 2024. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? *Advances in Neural Information Processing Systems*, 37:123846–123910.

Renguang Zuo, Oliver P Kreuzer, Jian Wang, Yihui Xiong, Zhenjie Zhang, and Ziye Wang. 2021. Uncertainties in gis-based mineral prospectivity mapping: Key types, potential impacts and possible solutions. *Natural Resources Research*, 30:3059–3079.

A STA-CoT

Algorithm 1 Reasoning Chain Construction

procedure BUILDCHAIN($Q, \mathcal{I}, K_{\text{domain}}, T_{\text{set}}$)
  $R^{(N)} \leftarrow \emptyset$; $M_{\text{cache}} \leftarrow \text{Initial}(Q, I)$
  $S^{(N)} \leftarrow M_p(Q, \mathcal{I}, K_{\text{domain}}, T_{\text{set}})$ ▷ Eq.3
  for $i = 1$ to N **do**
    $s^{(i)} \in S$; $ob^{(i)} \leftarrow M_e(s^{(i)})$ ▷ Eq.4
    $r^{(i)} \leftarrow (s^{(i)}, ob^{(i)})$ ▷ Form the reasoning node
    $R^{(i)} \leftarrow R^{(i-1)} \oplus R^{(i)}$ ▷ Eq.6
    $M_{\text{cache}} \leftarrow \text{Update}(r^{(i)})$
  end for
  return $R^{(N)}$
end procedure

Algorithm 2 Adaptive Stepwise Chain Refinement

Require: Question Q, Image $\mathcal{I}$, Domain Knowledge K_{domain}, Toolset T_{set}
Ensure: Final Answer A_{final}
Step 1: Reasoning Chain Construction
$A_{\text{candidate}} \leftarrow \emptyset$
$R^{(N)} \leftarrow$ BUILDCHAIN($Q, \mathcal{I}, K_{\text{domain}}, T_{\text{set}}$)
Step 2: Adaptive Stepwise Chain Refinement.
for $r = 1$ **to** K **do**
  $(isValid, A_p^{(r)}, F_{\text{corr}}) \leftarrow M_v(R^{(N)}, Q))$ ▷ Eq.5
  Add $A_p^{(r)}$ to $A_{\text{candidates}}$
  if $isValid$ **then**
    return $A_p^{(r)}$ ▷ Return validated answer
  end if
  for $x = 1$ **to** m **do**
    $ob_{re}^{(x_p)} \leftarrow M_e(s^{(x_p)}, f_{corr}^{(x_p)})$ ▷ Eq.4
    $r_{\text{re}}^{(x_p)} \leftarrow (s^{(x_p)}, ob_{\text{re}}^{(x_p)})$ ▷ Eq. 7
    $M_{\text{cache}} \leftarrow \text{Update}(M_{\text{cache}}, r_{\text{re}}^{(x_p)})$
  end for
  $R_{\text{re}}^{(N),r} \leftarrow R_{\text{prefix}} \oplus r_{\text{re}}^{(x_1)} \oplus \cdots \oplus r_{\text{re}}^{(x_m)}$ ▷ Rebuild
end for
Step 3: Fallback Global Majority Vote
if $A_{\text{candidates}} \neq \emptyset$ **then**
  return MAJORITYVOTE($A_{\text{candidates}}$) ▷ Eq.8
else
  return `"Failed"` ▷ All attempts failed
end if

B Case study

We present a detailed case study (Figure 6) to illustrate how STA-CoT effectively manages long, complex, and error-prone reasoning chains, addressing key challenges such as error propagation and global inconsistencies. Specifically, we highlight the coordinated roles of the rule-based controller and verifier in ensuring consistent multi-image geological reasoning. Initially, the reasoning process begins with Step 1, where the rule-based controller identifies a location mismatch error—discrepancy between the textual description and visual annotation. This detection immediately triggers a retry of Step 1, effectively preventing early-stage errors from propagating. Subsequent reasoning progresses until completion, at which point the verifier evaluates the entire chain and identifies two critical errors:

- **Step 2 error**: Incorrectly interpreting a grayscale image (M2) as containing colored, high-potential mineralization zones.
- **Step 5 error**: Faulty spatial relationship analysis derived directly from the incorrect Step 2 assessment.

These errors cause an initial validity check to fail (Valid Flag=0), necessitating targeted refinement.

In the refinement phase, STA-CoT triggers the refinement executor to address these specific errors:

- **Step 2 Retry**: Correctly recognizes image M2 as grayscale, accurately concluding no potential mineralization.
- **Step 5 Retry**: Revises spatial relationship analysis based on the corrected Step 2 outcome, recognizing insufficient spatial evidence to support a deposit.

Upon reevaluation, the verifier confirms that all prior errors have been successfully resolved (Valid Flag=1), and STA-CoT reaches a stable, consistent conclusion: *No deposit present*. This case exemplifies the robustness and reliability of STA-CoT, demonstrating how structured verification and targeted refinement effectively prevent error accumulation and ensure globally consistent reasoning outcomes.

C More Configuration Details

C.1 Expriment Setting

We evaluate a diverse set of open-source and closed-source multimodal large language models (MLLMs) on the MineBench dataset. For Qwen-VL-2.5-7B[2], we utilize official pretrained checkpoints and perform all local inference on two NVIDIA L40 GPUs (48GB each). For closed-source models such as GPT-4o-2024-08-06 and Gemini-2.0-Flash and Gemini-2.0-flash-thinking-exp-01-21, inference is conducted through their respective public APIs. When interacting with these

[2] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct

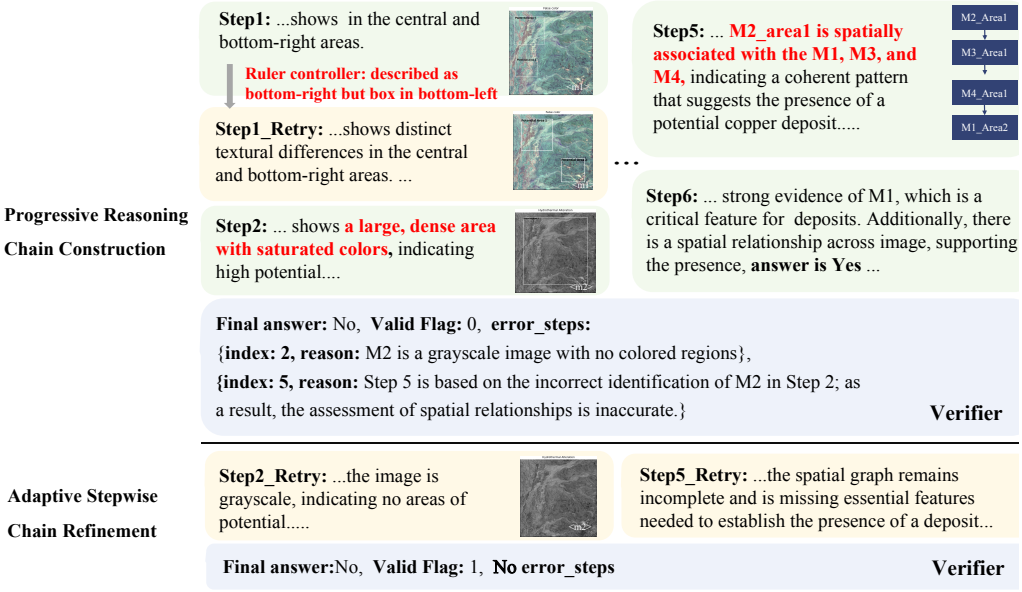


Figure 8: Visualization of qualitative example showcasing how our STA-CoT framework achieves successful, consistent multi-image geological reasoning.

APIs, the allocation of computational resources, memory, and execution time is fully managed by the service providers (OpenAI and Gemini). It is important to note that the Gemini series may be subject to automatic updates or model replacements by the provider, which can result in variations in performance over time. Specifically, Gemini-2.0 refers to the version released in February 2025. To ensure stability and comparability across all model outputs, we set the temperature to 0.05.

C.2 Spatial Graph Construction

To enable robust cross-image spatial reasoning, we construct a spatial graph that captures the geometric and relational structure among all candidate regions identified from multiple remote sensing images. The process begins by aggregating all bounding boxes (region proposals) extracted from each image. Each box is uniquely indexed and normalized to the respective image's coordinate system to ensure consistent comparison.

For every pair of images, we systematically analyze the relationships between all box pairs across the image set. Two primary spatial relationships are considered: *overlap* and *proximity*. Overlap relationships are established when the Intersection-over-Union (IoU) between two boxes from different images exceeds a predefined threshold, with the overlap degree used as a confidence score. If no overlap exists but the boxes are sufficiently close, we instead establish a proximity relationship, annotating the edge with a proximity-based confidence score. All spatial relationships are also tagged with their relative spatial direction (e.g., left, right, above, below, overlaps).

These detected relationships are used to construct a directed spatial graph, where each node corresponds to a candidate box, and edges encode the type and confidence of spatial relationships between regions from different images. This multi-image spatial graph is then aggregated and analyzed to identify connected components, groups of spatially coherent regions that may reflect geologically significant mineralization patterns spanning multiple images.

This spatial graph underpins subsequent post-hoc localization, area estimation, and multi-image geological reasoning steps, providing a structured and interpretable foundation.

C.3 Rule-Based Output Validation and Iterative Feedback

This appendix details our rule-based controller for automatic output validation and feedback-driven correction in the annotation process. The controller performs real-time checks on tool-generated bounding boxes and associated descriptions, immediately identifying common errors and issuing targeted prompts to guide annotation revision. Below, we summarize the primary error types and representative feedback provided by the system.

Region Size Errors. Marking excessively large regions can reduce specificity, while marking regions that are too small may capture irrelevant noise. Our controller automatically flags boxes that fall outside the acceptable size range and generates immediate feedback such as:

Feedback

1. Too large: avoid marking a large region. Please focus on specific features.
2. Too small: avoid marking small isolated spots and areas that are not clear evidence.

Overlapping or Redundant Regions Highly overlapping boxes are typically redundant and can hinder downstream analysis. The controller detects significant overlap (greater than 50%) and issues prompts such as:

Feedback

Some of your marked regions overlap significantly. Please remove redundant boxes and ensure each box covers a distinct, important feature.

Location-Description Mismatch To ensure interpretability and support spatial reasoning tasks, we cross-validate each box's coordinates with its textual description. If mismatches or coordinate system confusions are detected, the controller provides clear guidance, including a schematic reference of the coordinate system:

Feedback

There is confusion between 'left' and 'right'. In our coordinate system, x=0 corresponds to the LEFT edge and x=1 to the RIGHT edge of the image. Please double-check that your spatial description (e.g., "top-left") matches the actual box position in [ymin, xmin, ymax, xmax] format. Reference:

```
(0,0) ----> (0,1)
  |           |
  v           v
(1,0) ----> (1,1)
```

Excessive Total Area or Region Count To prevent loss of focus and annotation noise, the controller enforces limits on both total marked area and region count. When these limits are exceeded, the system provides feedback such as:

Feedback

You have marked too many regions or the total area covered is excessive. Please limit your annotation to at most five regions and ensure the combined area covers less than 70% of the image.

Iterative Feedback and Correction For each detected error, the STA-CoT generates feedback that evolves over multiple correction rounds. Initial feedback is general and constructive, while repeated violations elicit increasingly detailed, operational guidance. On the final attempt, feedback includes explicit instructions and a coordinate system reference to resolve any remaining ambiguities, as following:

Feedback

This is your final attempt. Please ensure:
1. All regions are appropriately sized (neither too large nor too small).
2. Location descriptions match actual box coordinates ([ymin, xmin, ymax, xmax], with (0,0) at the top-left).
3. If no salient features are present, you may leave the coordinates empty and state this in your description.
Reference:

```
(0,0) ----> (0,1)
  |           |
  v           v
(1,0) ----> (1,1)
```

This error-driven, feedback-enhanced controller forms the backbone of our automated annotation quality assurance, ensuring robust, interpretable, and high-quality region outputs for vision-grounded Executor.

C.4 Detailed Prompt Construction for STA-CoT

We provide detailed prompts for each module and the toolkit of STA-CoT.

Planner

You are an expert in remote sensing and task planning. Your task is to analyze the following question and create a detailed, logically ordered plan.

Provided Information

- **Question**: [The analysis question]
- **Domain Knowledge Base**: [Relevant context, if any]
- **Available Tools**: [List of suggested tools]

Task Guidelines

1. **Task Decomposition:**
    - Break down the analysis into clear, sequential steps.
    - Each step should be self-contained and focused.
    - Steps should follow a logical progression.
    - Include clear expected outcomes.
2. **Resource Selection:**
    - Choose **ONLY** the most relevant images for each step.
    - Explain why each image is needed.
3. **Tool Selection:**
    - Only include tools when visual analysis is necessary.
    - Each tool must have a clear purpose.
    - Explain why each tool is needed.

Result Format

```
[
  {
    "Step": "Step number and name",
    "Thought": "Why this step is necessary and what it provides",
    "Action": [
      {
        "Suggested Tool": "Tool name (if needed)",
        "Action": "Specific action and why this tool is needed"
      }
    ],
    "Resources": ["List of specific images needed for this step"]
  }
]
```

Executor

You are a remote sensing expert specializing in mineral exploration. Your task is to execute a specific analysis step using the provided images and tools.

Provided Information

- **Current Step**: [Description of current subtask]
- **Previous Analysis**: [Prior step outputs, if any]
- **Available Tools**: [List of suggested tools]

Task Guidelines

1. **Image Assessment:**
    - Carefully examine the provided image(s).
    - Identify features or patterns relevant to the subtask.
    - If no clear features are found, answer "No" without using tools.
    - If features are found, proceed with tool usage.
2. **Tool Selection and Usage:**
    - Use tools **ONLY** if necessary for the current step.
    - Return the tool parameters.
3. **Evidence Collection:**
    - Document all visual evidence found.
    - Explain how the evidence supports or contradicts the analysis.
    - If using multiple images, explain the relationships between them.

Result Format

```
[
    {
        "Step": "Current step number and description",
        "Tools": [
            {
                "Name": "Tool name",
                "Purpose": "Specific reason for using this tool",
                "Parameters": "The parameters of tool"
            }
        ],
        "Result": {
            "Sub_Inference": "Yes|No",
            "Explanation": "Detailed findings and evidence",
            "Global_Inference": "end|Unknown"
        }
    }
]
```

Result Format Guidelines

- **Sub_Inference**: "Yes" only with strong and clear visual evidence, otherwise "No".
- **Explanation**: Clearly connect evidence to conclusions.
- **Global_Inference**: "end" if concluding, "Unknown" if continuing.

Verifier

You are an expert evaluator tasked with assessing the **correctness, coherence, and sufficiency** of a multi-step reasoning process that uses visual tools to answer geoscientific questions.

Provided Information

- **Question**: [Task question]
- **Reasoning Process**: [Prior knowledge/reasoning steps]
- **Available Tools**: [List of visual tools]

Evaluation Criteria

1. **Reasoning Structure & Validity**
    - Is the reasoning logically structured and complete?
    - Are all necessary steps present, without major gaps or unsupported claims?
    - Minor flaws are acceptable only if they do not undermine the overall conclusion.
2. **Tool Usage & Interpretation**
    - Were the appropriate tools selected and applied correctly?
    - Are tool outputs interpreted accurately and relevant to the final conclusion?
    - Tool use is optional only if direct visual analysis clearly provides visual evidence.
3. **Answer Justification**
    - **Conclude *yes*** only if all **CRITICAL** features are clearly supported and **SECONDARY** features are present.
    - **Conclude *no*** if any **CRITICAL** feature lacks evidence, contradictions or tool errors undermine CRITICAL claims, or a CRITICAL feature is missing at any step.

Valid Flag

- **Score 1 (Valid Reasoning):** Logically sound, accurate, well-supported; appropriate tool usage; final answer matches tool outputs.
- **Score 0 (Flawed Reasoning):** Missing/incorrect CRITICAL evidence; tool misuse; misinterpretation; contradictions; unsupported CRITICAL features.

Efficiency Rule

- If a decisive flaw is found (e.g., unsupported CRITICAL feature): Skip remaining steps; score as 0 with justification.

Apply efficiency logic to minimize unnecessary re-analysis once a conclusive flaw is identified.

Result Format

```
[
  {
    "evaluation": {
      "score": "1 or 0",
      "assessment": {
        "final_answer": "yes or no",
        "main_reasoning": "Summary of key reasoning steps and supporting evidence",
        "tool_evaluation": "Evaluation of tool usage and interpretation",
        "reasoning_evaluation": "Assessment of logical structure, completeness, and integration of
            tool outputs"
      },
      "feedback": {
        "issues": "List of reasoning flaws if score is 0 (e.g., missing steps, tool misuse,
            misinterpretation, contradictions)",
        "error_steps": [
          { "index": 1, "reason": "Misinterpretation of tool output" },
          { "index": 2, "reason": "Incorrectly highlighted area in tool output" }
        ]
      }
    }
  }
]
```

Tool Definitions

Available tools are defined as follows:

```
{
  "box_maker_color_mode": {
    "description": "Identify and highlight continuous color regions based on visual color
        distribution. Focus exclusively on color regions aligned with the color bar, ignoring the
        grayscale background.",
    "parameters": {
      "Coordinates": "List of bounding box coordinates in normalized form: [[ymin, xmin, ymax, xmax],
          ...]",
      "Image Resource": "List of image filenames used for color analysis",
      "Description": "Explain the significance of each marked region and describe its position (e.g.,
          'top-left corner shows a large, dense yellow-orange patch indicating high potential'). If
          none are found, provide reasoning",
      "Location": "List of approximate locations corresponding to each coordinate set (e.g., ['top-
          left', 'center-right', 'bottom-center', ...]) to verify coordinates match described
          positions"
    }
  },

  "box_maker_feature_region": {
    "description": "Mark geologically or structurally significant regions, including patterns,
        textures, or anomalies related to geological formations, structural features, or mineral
        deposits.",
    "parameters": {
      "Coordinates": "List of bounding box coordinates in normalized form: [[ymin, xmin, ymax, xmax],
          ...]",
      "Image Resource": "List of image filenames used for analysis",
      "Description": "Explain each marked region's geological significance and position. If none are
          found, provide reasoning.",
      "Location": "List of approximate locations corresponding to each coordinate set (e.g., ['top-
          left', 'center-right', 'bottom-center', ...]) to verify coordinates match described
          positions"
    }
  },

  "spatial_relationship_explorer": {
    "description": "Analyze the spatial relationships between marked regions in the provided images.
        Step 1: Marking verification (label each as 'CORRECT' or 'INCORRECT'). Step 2: Analyze
        spatial relationships among correctly marked regions.",
    "parameters": {
      "Image Resource": "List of image filenames used",
      "marking_accuracy_verification": "A list labeling each region as 'CORRECT' or 'INCORRECT', with
          reasoning",
      "spatial_relationships": "List of spatial relationships identified among correctly marked
          regions, including type, description, and justification"
    }
  },

  "decision_making": {
    "description": "Determine the final answer by evaluating both the color-based potential of marked
        areas and their spatial relationships, with emphasis on whether critical areas show at
        least moderate potential.",
    "parameters": {
      "Image Resource": "List of image filenames used",
      "critical_areas_assessment": "Whether all critical areas show at least moderate potential and
          are clearly visible",
      "relationship_validity": "Whether spatial relationships between critical areas are accurate and
          visually supported"
    }
  }
}
```

Localization Tool

You are a professional geological spatial analysis expert specializing in analyzing mineral alteration zones using image information and spatial relationships. Please help determine the most appropriate parameter settings for running the geological relationship analysis algorithm.

Optimization Goals:

1. **High Recall:** Ensure all truly important geological areas are captured in the prediction (minimize false negatives).
2. **High Area Reduction:** Minimize the total predicted area while maintaining high recall (maximize efficiency).

Algorithm Core Parameters:

- **Centrality metric weights:**
  - **degree_centrality_weight** [RANGE: 1.0–3.0, DEFAULT: 1.0] – Measures how many direct connections an alteration zone has to other zones.
  - **betweenness_centrality_weight** [RANGE: 1.0–3.0, DEFAULT: 1.0] – Identifies zones that serve as bridges between different geological features.
- **Chain-related weights:**
  - **chain_presence_weight** [RANGE: 1.0–3.0, DEFAULT: 1.0] – Emphasizes intersection zones where multiple geological processes overlap.
- **Area type weights (determine from dialogue):**
  - **Silicification Zone** [RANGE: 1.0–2.0, DEFAULT: 1.0]
  - **Propylitic Alteration** [RANGE: 1.0–2.0, DEFAULT: 1.0]
  - **Hydrothermal Alteration** [RANGE: 1.0–2.0, DEFAULT: 1.0]
  - **False Color Composition** [RANGE: 1.0–2.0, DEFAULT: 1.0]
    (Different alteration types have varying associations with mineral deposits.)
- **Boundary parameters:**
  - **important_areas_count** [RANGE: 3–7, INTEGER] – Controls how many high-priority areas the algorithm will focus on.
  - **hop2_expansion_height** [RANGE: 1.0–1.5] – Vertical expansion multiplier for hop2 level, to capture peripheral mineralization.
  - **hop2_expansion_width** [RANGE: 1.0–1.5] – Horizontal expansion multiplier for hop2 level, to capture peripheral mineralization.

Note: Minerals typically have radiation zones extending beyond their central concentrations. These expansion parameters are crucial to avoid missing peripheral mineralization at the edges of alteration zones and to capture transitional boundaries where valuable deposits may exist.

Resutl Format:

```
{
  "degree_centrality_weight": "float, range 1.0-3.0, default 1.0",
  "betweenness_centrality_weight": "float, range 1.0-3.0, default 1.0",
  "chain_presence_weight": "float, range 1.0-3.0, default 1.0",
  "area_type_weights": {
    "Silicification Zone": "float, range 1.0-2.0, default 1.0",
    "Propylitic Alteration": "float, range 1.0-2.0, default 1.0",
    "Hydrothermal Alteration": "float, range 1.0-2.0, default 1.0",
    "False Color Composition": "float, range 1.0-2.0, default 1.0"
  },
  "important_areas_count": "integer, range 3-7",
  "hop2_expansion_height": "float, range 1.0-1.5",
  "hop2_expansion_width": "float, range 1.0-1.5"
}
```