AreoRAG: Hyperbolic Spatial Hypergraph and Physics-Informed Conflict Triage for Multi-Source Planetary Retrieval Augmented Generation

Ao Long, Ze Deng, and Lizhe Wang†

***Abstract*—Retrieval Augmented Generation (RAG) has demonstrated considerable promise in grounding Large Language Models (LLMs) with external knowledge for knowledge-intensive question answering. However, extending RAG to the domain of planetary science — where multi-source remote sensing observations are inherently embedded in continuous physical space and inter-source disagreements often carry scientific value — introduces fundamental challenges that existing multi-source RAG frameworks cannot address. These challenges manifest in two critical aspects: (1) existing discrete graph topologies (e.g., multi-source line graphs) suffer from edge explosion when encoding continuous spatial proximity, failing to bridge the gap between physical continuity and semantic discreteness; and (2) conventional conflict-filtering mechanisms, designed under the assumption that inter-source inconsistency implies unreliability, systematically suppress scientifically valuable observational disagreements that are intrinsic to multi-platform deep-space exploration. To address these challenges, we propose AreoRAG, a novel framework tailored for multi-source planetary spatial data retrieval augmented generation. Our framework introduces two key innovations: (1) a Hyperbolic Spatial Hypergraph (HySH) construction module that employs n-ary spatial observation hyperedges embedded in hyperbolic space via the Lorentz model, where spatial resolution is coupled with radial depth to faithfully represent the hierarchical scale structure of planetary observations while reducing edge complexity from $O(k^2)$ to $O(k)$; and (2) a Physics-Informed Conflict Triage (PICT) module that detects inter-source conflicts via cross-source interaction entropy, classifies them into four physically grounded categories (noise, instrument-inherent, scale-dependent, and temporal-evolution), and applies differentiated confidence recalibration to preserve scientifically valuable disagreements while filtering genuine noise. Extensive experiments on multi-source planetary observation datasets demonstrate that AreoRAG significantly enhances both the retrieval fidelity and the scientific faithfulness of knowledge-augmented generation in planetary science scenarios.**

***Index Terms*—Retrieval Augmented Generation, Planetary Remote Sensing, Hypergraph, Hyperbolic Space, Knowledge Conflict Triage, Multi-source Spatial Data, Mars Exploration.**

I. Introduction

THE past two decades have witnessed an unprecedented accumulation of multi-source remote sensing data from Mars exploration missions. Orbital platforms, such as the Mars Reconnaissance Orbiter, Mars Express, and Tianwen-1, continuously acquire observations across diverse modalities. These modalities range from sub-meter optical imagery (HiRISE) [1] and medium-resolution contextual mosaics (CTX) [2] to hyperspectral mineralogical mapping (CRISM) [3] and global topographic models (MOLA) [4]. Simultaneously, surface assets including the Curiosity [5] and Zhurong rovers [6] generate complementary in-situ measurements through spectrometers, ground-penetrating radar, and navigation cameras. This rapidly expanding, multi-source, multi-resolution data ecosystem has created a pressing demand for intelligent knowledge retrieval systems that can support planetary scientists in conducting semantic search, cross-source correlation, and multi-scale reasoning over heterogeneous observation archives [7].

Large Language Models (LLMs) have emerged as powerful tools for natural language understanding and generation [8], and Retrieval Augmented Generation (RAG) has been established as a standard paradigm for grounding LLM responses in external knowledge bases [9]. By dynamically retrieving relevant documents and conditioning generation on retrieved context, RAG effectively mitigates the hallucination problem inherent in LLMs and enables knowledge-intensive question answering [10]. The synergy between LLMs and Knowledge Graphs (KGs) has further advanced retrieval performance through structured knowledge representation, achieving notable improvements in multi-hop reasoning, credibility assessment, and interpretability [11].

Nevertheless, deploying RAG systems for planetary science knowledge retrieval introduces domain-specific complexities that fundamentally challenge existing frameworks. Recent advances in multi-source RAG, exemplified by MultiRAG [12], have made significant progress through multi-source line graphs and multi-level confidence computation. However, when confronted with planetary spatial data, these methods encounter two fundamental problems that cannot be resolved through parameter tuning alone:

1) **The Spatial Topology Loss Problem.** Conventional multi-source retrieval systems judge relevance by textual semantic similarity. Planetary observations are different. Each observation is tied to a spatial footprint on the surface, a time window, and a set of instrument parameters. Two observations are relevant to each other mainly because they are spatially close, temporally overlap-

A. Long, W. Lin and Z. Deng, (Corresponding author, dengze@cug.edu.cn) are with the School of Computer Science, China University of Geosciences, Wuhan, 430078, P. R. China.

Z. Deng is also with Hubei Key Laboratory of Intelligent Geo-Information Processing, China University of Geosciences, Wuhan 430074, China.

ping, or captured at complementary resolutions. Existing methods such as multi-source line graphs [12] build graph topology from discrete text entities. This design creates a mismatch with continuous spatial data: k co-located entities need $\binom{k}{2} = O(k^2)$ pairwise edges to represent their spatial relationships. The resulting edge explosion removes the sparsity that these graph models rely on. In short, the discrete graph structure cannot bridge the gap between physical continuity and semantic discreteness.

2) **The Conflict Over-Smoothing Problem.** Existing multi-source RAG frameworks treat inter-source inconsistency as misinformation or hallucination. They use confidence scores to remove conflicting nodes [12], [13]. In planetary science, however, different platforms naturally produce different measurements for the same target. An orbiter and a rover observe at different scales, depths, and wavelengths. For example, an orbital spectrometer may detect hydrated minerals on the surface, while an in-situ drill finds olivine-carbonate assemblages below. This conflict does not come from data error. It reflects geological evolution across depth. If we apply uniform conflict filtering, the system suppresses these scientifically valuable signals together with genuine noise. This over-smoothing violates a core principle of deep-space exploration: observational disagreements should be preserved, because they may lead to new discoveries through multi-source comparison.

To address these two challenges, we propose AreoRAG, a framework designed for multi-source planetary spatial data retrieval augmented generation. AreoRAG introduces two innovations. We first construct a **Hyperbolic Spatial Hypergraph (HySH)** to resolve the spatial topology loss problem. HySH uses n-ary spatial observation hyperedges to group co-located multi-source observations into single high-order facts. This design reduces edge complexity from $O(k^2)$ to $O(k)$. We embed these hyperedges in hyperbolic space via the Lorentz model. The exponential volume growth of negative-curvature geometry naturally fits the hierarchical scale structure of planetary observations. Coarse-resolution global data resides near the origin, while fine-resolution local data extends toward the boundary. To resolve the conflict over-smoothing problem, we develop a **Physics-Informed Conflict Triage (PICT)** mechanism. PICT replaces uniform conflict filtering with a differentiated triage strategy. It first detects inter-source conflicts through cross-source interaction entropy. Then it classifies each conflict into one of four physically grounded categories: noise, instrument-inherent, scale-dependent, and temporal-evolution. Finally, it applies category-specific confidence recalibration, filtering genuine noise while provably preserving scientifically valuable observational disagreements. The two modules form a tightly coupled loop. HySH provides spatially faithful multi-source evidence to PICT, while PICT feeds back triage results to prioritize scientifically interesting regions in subsequent retrieval.

The contributions of this paper are summarized as follows:

1) We propose a Hyperbolic Spatial Hypergraph (HySH) construction module for multi-source planetary data, by combining the n-ary hyperedge representation from hypergraph-based RAG [**?**] with the Lorentz-model hyperbolic embedding from hyperbolic knowledge graph methods [**?**]. HySH couples spatial resolution with hyperbolic radial depth so that the hierarchical scale structure of planetary observations is preserved, while edge complexity is reduced from $O(k^2)$ to $O(k)$. We further propose a resolution-aware Spatial Outward Einstein Midpoint (Spatial OEM) aggregation operator with a formal guarantee of outward bias.
2) We propose a Physics-Informed Conflict Triage (PICT) mechanism for multi-source retrieval, by adapting the entropy-based conflict detection from [**?**] and the linear-separability finding of knowledge conflicts from [**?**]. PICT classifies each inter-source conflict into four physically grounded categories (noise, instrument-inherent, scale-dependent, temporal-evolution) and applies category-specific confidence recalibration. We provide a formal Anti-Over-Smoothing Guarantee showing that scientifically valuable disagreements are provably preserved. To the best of our knowledge, this is the first conflict-handling mechanism in RAG that explicitly distinguishes erroneous inconsistency from scientifically meaningful observational divergence.
3) We design the AreoRAG Prompting (ARP) algorithm that integrates HySH and PICT through three coupling points: spatial alignment as a prerequisite for interaction entropy computation, radial depth difference as a resolution disparity signal for conflict classification, and triage-driven retrieval priority feedback. Experiments on three Mars observation datasets show that AreoRAG outperforms existing multi-source RAG methods in both retrieval accuracy and conflict preservation.

II. Preliminary

In the field of planetary spatial knowledge retrieval, the primary challenges include faithfully representing continuous spatiotemporal relationships across heterogeneous observation sources and achieving reliable retrieval under inherent inter-source scientific conflicts. This section introduces the core elements of our approach and precisely defines the problems we address.

Let $Q = \{q_1, q_2, \ldots, q_n\}$ be the set of query instances, where each q_i corresponds to a distinct planetary science query. Let $\mathcal{E} = \{e_1, e_2, \ldots, e_m\}$ be the set of entities in the spatial knowledge hypergraph, where each e_j represents a geological feature, instrument, or observation product. Let $\mathcal{R} = \{r_1, r_2, \ldots, r_p\}$ be the set of relationships, and let $\mathcal{F} = \{f_1^n, f_2^n, \ldots, f_s^n\}$ be the set of n-ary relational facts (hyperedges). Let $D = \{d_1, d_2, \ldots, d_t\}$ be the set of observation data products, where each d_l represents an observation record from a specific instrument. We define the spatially-grounded knowledge-guided retrieval augmented generation problem as follows:

$$\arg\max_{d_i \in D} \mathrm{LLM}\left(q_i, d_i\right), \tag{1}$$

$$\sum_{e_j \in \mathcal{E}} \sum_{f_k^n \in \mathcal{F}} \mathrm{HG}\left(e_j, f_k^n, d_i\right) \cdot \mathcal{S}_{geo}\left(q_i, d_i\right), \tag{2}$$

where $\mathrm{LLM}(q_i, d_l)$ denotes the relevance score between query q_i and document d_l assessed by the LLM, $\mathrm{HG}(e_j, f_k^n, d_l)$ represents the degree of match between entity e_j, n-ary fact f_k^n, and document d_l in the hypergraph, and $\mathcal{S}_{geo}(q_i, d_i)$ is a spatial compatibility function that ensures the retrieved evidence satisfies the geospatial constraints (footprint overlap, temporal window, resolution range) specified in the query.

Furthermore, we optimize the knowledge construction and retrieval modules by introducing a hyperbolic spatial hypergraph to achieve spatially faithful knowledge aggregation and physics-informed conflict handling. Specifically, the proposed approach is formally defined through the following definitions.

Definition 1 (Multi-source planetary observation data). Given a set of observation platforms $\mathcal{H}$ (e.g., MRO, Mars Express, Tianwen-1, Curiosity, Zhurong), the observation data $D = \{\mathcal{I}, \mathcal{P}_{foot}, \mathcal{T}_{win}, \mathcal{S}_{band}, c, \mathrm{meta}\}$ exists, where $\mathcal{I}$ denotes the instrument identity, $\mathcal{P}_{foot} \subset \mathbb{S}^2_{Mars}$ denotes the spatial footprint on the Martian surface, $\mathcal{T}_{win}$ denotes the temporal acquisition window parameterized by Solar Longitude L_s, $\mathcal{S}_{band}$ denotes the spectral band configuration, c represents the observation content (image, spectrum, or derived product), and meta represents the PDS/CNSA metadata. Through a multi-source spatial adapter parsing algorithm, we obtain normalized data $\widehat{D} = \{\mathrm{id}, \mathcal{I}, \mathcal{P}_{foot}, \mathcal{T}_{win}, \mathcal{S}_{band}, \ell_{res}, \mathrm{jsc}, \mathrm{meta}\}$, where id is the unique identifier, $\ell_{res} \in \mathbb{R}^+$ denotes the ground sampling distance (spatial resolution), and jsc denotes the observation content stored using JSON-LD for linked data interoperability.

Definition 2 (N-ary spatial knowledge hypergraph). An n-ary spatial knowledge hypergraph is defined as $\mathcal{G}_{hyp} = (\mathcal{E}, \mathcal{R}, \mathcal{F}_{spa})$, where $\mathcal{E}$ denotes the entity set, $\mathcal{R}$ denotes the relation set, and $\mathcal{F}_{spa}$ denotes the set of spatial observation hyperedges. Each spatial observation hyperedge $f_{spa}^n \in \mathcal{F}_{spa}$ binds multiple entities and observation parameters into a single n-ary relational fact:

$$f_{spa}^n = \left(\mathcal{I},\ \mathcal{P}_{foot},\ \mathcal{T}_{win},\ \mathcal{S}_{band},\ \mathcal{O}_{target},\ \ell_{res}\right), \tag{3}$$

where $\mathcal{O}_{target}$ denotes the set of target geological features. Unlike binary knowledge graphs where k co-located entities require $\binom{k}{2} = O(k^2)$ pairwise edges, a single n-ary hyperedge binds all k entities with $O(k)$ complexity, directly resolving the edge explosion problem.

Definition 3 (Hyperbolic space embedding). We represent $\mathcal{G}_{hyp}$ in d-dimensional hyperbolic space $\mathbb{H}^d_K$ with constant negative curvature $K < 0$ using the Lorentz (hyperboloid) model. The hyperbolic space is realized as:

$$\mathbb{H}^d_K = \left\{ \mathbf{x} \in \mathbb{R}^{d+1} \mid \langle \mathbf{x}, \mathbf{x} \rangle_L = \frac{1}{K},\ x_0 > 0 \right\} \tag{4}$$

where $\langle \mathbf{x}, \mathbf{y} \rangle_L = -x_0 y_0 + \sum_{i=1}^d x_i y_i$ is the Lorentzian inner product. The geodesic distance between two points $\mathbf{x}, \mathbf{y} \in \mathbb{H}^d_K$ is $d_K(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{-K}} \cosh^{-1}(K \langle \mathbf{x}, \mathbf{y} \rangle_L)$. The radial depth $r(\mathbf{x}) = x_0$ encodes the intrinsic distance from the origin and serves as a proxy for hierarchical specificity: entities near the origin represent coarse, global-scale features, while those at large radial depth represent fine-scale, local observations.

Definition 4 (Observation-grounded homologous data). For a query $Q(q, \mathcal{G}_{hyp})$ on the spatial hypergraph $\mathcal{G}_{hyp}$, the multi-source spatial evidence retrieved in a single query is defined as observation-grounded homologous data. For any two observations v_1 and v_2 in $\mathcal{G}_{hyp}$, they are observation-grounded homologous if and only if they: (a) belong to the same retrieval candidate set, and (b) their spatial footprints satisfy $\mathcal{P}_{foot}(v_1) \cap \mathcal{P}_{foot}(v_2) \neq \varnothing$.

Definition 5 (Observation-grounded knowledge source). A planetary observation knowledge source is defined as $\mathcal{K}_s = (\mathcal{I}_s, \Omega_s, F(\mathcal{K}_s), \mathcal{M}_s)$, where $\mathcal{I}_s$ denotes the instrument, $\Omega_s = (\ell_{res}, \lambda_{band}, \theta_{view}, d_{pen})$ denotes the observation geometry parameters (spatial resolution, spectral band, viewing angle, penetration depth), $F(\mathcal{K}_s)$ denotes the set of atomic factual statements, and $\mathcal{M}_s$ denotes the physical measurement model that maps target properties through observation constraints to observable facts.

Definition 6 (Conflict triage confidence.) For observation-grounded homologous data obtained from the spatial hypergraph, the conflict triage confidence integrates two levels of assessment: (a) cross-source interaction entropy to detect inter-source conflicts, and (b) physics-informed conflict classification to determine whether detected conflicts represent noise to be filtered or scientifically meaningful observational divergences to be preserved. Unlike conventional candidate confidence [14] that uniformly penalizes inconsistency, conflict triage confidence applies differentiated recalibration based on the physical origin of each conflict.

III. Methodology

A. Framework of AreoRAG

This section elaborates on the implementation approach of AreoRAG. As shown in Fig. 3, the framework comprises three tightly coupled modules. The first step involves constructing a Hyperbolic Spatial Hypergraph (HySH) from multi-source planetary observation data, achieving unified spatiotemporal representation via n-ary observation hyperedges embedded in hyperbolic space (Section III-B); the second step performs spatiotemporal retrieval on the constructed HySH, where hyperbolic spatial proximity encoding and cross-resolution aggregation via the Spatial Outward Einstein Midpoint are employed to extract query-relevant multi-source evidence (Section III-C); the third step applies Physics-Informed Conflict Triage (PICT), which detects inter-source conflicts via cross-source interaction entropy, classifies them into four scientific categories, and executes conflict-aware confidence recalibration to preserve scientifically valuable disagreements while filtering noise (Section III-D). Finally, integrating the aforementioned steps to form the AreoRAG Prompting algorithm, ARP (Section III-E).

The three modules interact through three explicit coupling points: (1) HySH's spatial alignment is a prerequisite for meaningful interaction entropy computation in PICT; (2) the radial depth difference Δr from HySH directly feeds into the

PICT feature vector as the resolution disparity signal; and (3) PICT's triage results feed back to boost retrieval priority of scientifically interesting regions in subsequent queries.

B. Hyperbolic Spatial Hypergraph Construction

The AreoRAG method begins by constructing a knowledge structure that can faithfully represent the continuous spatiotemporal topology of planetary multi-source data. Unlike MultiRAG's Multi-source Line Graph (MLG), which relies on discrete text entities and binary triples, we introduce a hypergraph structure embedded in hyperbolic space to jointly address edge explosion and spatial scale hierarchy.

1) Multi-source Spatial Adapter Parsing: We first design a spatial adapter for each observation data source to parse instrument metadata, spatial footprints, temporal windows, and spectral parameters. For orbital remote sensing data (e.g., HiRISE, CTX, CRISM), parsing involves extracting the image footprint geometry, ground sampling distance, and spectral band configuration from PDS labels. For in-situ data (e.g., rover spectrometers, ground-penetrating radar), parsing extracts the rover traverse coordinates, measurement timestamps in Sol, and instrument-specific parameters such as penetration depth. All temporal references are unified to Solar Longitude L_s to enable cross-platform temporal comparison. For derived data products (e.g., DTMs, mineral abundance maps), parsing extracts provenance links to the source observations and processing parameters.

The final integration of multi-source spatial data can be expressed as:

$$D_{Fusion} = \bigcup_{i=1}^{n} A_i^{spa}(D_i), \tag{5}$$

where $A_i^{spa} \in \{Ada_{orbital}, Ada_{insitu}, Ada_{derived}\}$ represents the spatial adapter parsing functions for orbital, in-situ, and derived data products respectively, and D_i represents the original observation datasets from different platforms.

Through the parsed data D_{Fusion}, we further extract entities (geological features, mineral signatures, topographic structures), relationships (spatial containment, temporal succession, compositional association), and observation-specific attributes. The knowledge extraction process employs LLM-based entity recognition guided by a planetary science domain schema:

$$\sum_{D_i} \left(\{e_1, e_2, \ldots, e_m\} \sqcup \{r_1, r_2, \ldots, r_n\} \sqcup \{f^n_{spa,1}, \ldots, f^n_{spa,p}\} \right). \tag{6}$$

2) Spatial Observation Hyperedge Formation: Based on the extracted knowledge base, we construct spatial observation hyperedges that bind co-located multi-source observations into single n-ary facts. As formalized in Definition 2, each hyperedge f^n_{spa} encapsulates the instrument, spatial footprint, temporal window, spectral bands, target features, and resolution. In a pairwise binary graph, k co-existing spatial entities require $\binom{k}{2} = O(k^2)$ spatial proximity edges. With hyperedges, a single n-ary fact binds all k entities, reducing edge complexity to $O(k)$. This directly resolves the edge explosion problem identified in our analysis of MLG.

3) Scale-Aware Lorentz Embedding: We embed the spatial observation hypergraph in d-dimensional hyperbolic space $\mathbb{H}^d_K$ using the Lorentz model (Definition 3). The key innovation is coupling the radial depth with spatial resolution through an embedding mapping $\Phi : \mathcal{F}_{spa} \to \mathbb{H}^d_K$:

$$r\left(\Phi(f^n_{spa})\right) = \frac{1}{\sqrt{-K}} \cosh\left(\sqrt{-K} \cdot g(\ell_{res})\right), \tag{7}$$

where $g(\ell_{res}) = -\log(\ell_{res}/\ell_{max})$ is a monotone decreasing function of resolution, and $r(\mathbf{x}) = x_0$ denotes the radial depth.

This embedding design is motivated by the following observation on the intrinsic geometry of planetary spatial data:

Proposition 1 (Spatial Scale-Curvature Correspondence). *The planetary spatial observation hierarchy exhibits tree-like branching: each coarser-resolution observation spatially contains multiple finer-resolution observations. Let $N(\ell)$ denote the number of observations at resolution level ℓ. For remote sensing data with total survey area $A_{coverage}$:

$$N(\ell) \propto A_{coverage}/\ell^2. \tag{8}$$

As resolution ℓ decreases (finer scale), $N(\ell)$ grows quadratically, exhibiting the exponential branching characteristic of negative-curvature spaces. Therefore, the spatial scale hierarchy is intrinsically hyperbolic, and Euclidean embedding with polynomial volume growth cannot faithfully represent it.*

Through this embedding, global coarse-resolution data (e.g., MOLA topography at 460 m) is placed near the hyperbolic origin (small radial depth), while local high-resolution data (e.g., HiRISE at 0.3 m) is placed far from the origin (large radial depth). The exponential volume growth of $\mathbb{H}^d_K$ naturally accommodates the exponentially increasing number of observations at finer scales.

4) Cross-Reference-Frame Alignment: To address the heterogeneous reference frame problem (orbiter areocentric coordinates vs. rover-centric local coordinates), we align all observations to a global reference via parallel transport on the hyperbolic manifold:

$$\Phi_{aligned}(e) = \exp_{o_g}\left(\Gamma_{o_k \to o_g}\left(\log_{o_k}(\Phi_k(e))\right)\right), \tag{9}$$

where $\log_{o_k}$ is the logarithmic map at the local reference origin o_k, $\Gamma_{o_k \to o_g}$ is the parallel transport operator along the geodesic from o_k to the global origin o_g, and $\exp_{o_g}$ is the exponential map at the global origin. Unlike Euclidean affine transformations, hyperbolic parallel transport preserves geodesic distances and radial depth, ensuring that scale hierarchy information is maintained after cross-frame alignment.

Here, we provide a simple example of hyperbolic spatial hypergraph construction. As shown in Fig. 4, an observation region is covered by three sources at different resolutions: a CTX mosaic (6 m), an HiRISE strip (0.3 m), and a CRISM spectral cube (18 m). In the HySH, the HiRISE observation (finest resolution) is embedded at the largest radial depth, while the CRISM observation (coarsest resolution) is nearest to the origin. A spatial observation hyperedge binds all three observations and their co-located geological features into a single n-ary fact, without requiring $O(k^2)$ pairwise edges.

C. Spatiotemporal Retrieval with Cross-Resolution Aggregation

After the construction of the hyperbolic spatial hypergraph, the next step is to retrieve query-relevant multi-source spatial evidence. The retrieval process comprises two phases: spatiotemporal evidence extraction and cross-resolution aggregation.

1) Spatial Intent Extraction and Hyperedge Retrieval: Given a user query q, we first employ the LLM to extract spatial intent, including target entities, spatial constraints (footprint, region), temporal constraints (L_s range, Sol range), and resolution preferences. These are denoted as query elements $\mathcal{K}_q$.

For each topic entity $e_s \in \mathcal{E}_q$ extracted from the query, we retrieve its incident spatial observation hyperedges $\mathcal{F}_{e_s} = \{f^n_{spa} \in \mathcal{F}_{spa} : e_s \in f^n_{spa}\}$ and derive pseudo-binary triples (e_h, f^n_{spa}, e_t) for pairwise reasoning, following the approach of HyperRAG [18]:

$$\mathcal{T}_q = \left\{(e_h, f^n_{spa}, e_t) \mid f^n_{spa} \in \mathcal{F}_{e_s},\ e_h \in f^n_{spa},\ e_t \in f^n_{spa}\right\}. \tag{10}$$

2) Hyperbolic Spatial Encoding and Plausibility Scoring: For each candidate triple, we compute a spatiotemporal encoding that fuses semantic, structural, and physical-spatial signals:

$$\mathbf{x} = \left[\varphi(q) \| \varphi(e_h) \| \varphi(f^n_{spa}) \| \varphi(e_t) \| \delta(e_h, f^n_{spa}, e_t) \| \psi_{geo}(e_h, e_t)\right], \tag{11}$$

where φ denotes a text embedding model, δ denotes a structural proximity encoding adapted from SubGraphRAG [19] to operate on hyperedges, and ψ_{geo} is the hyperbolic spatial encoding defined as:

$$\psi_{geo}(e_h, e_t) = \left[d_K\left(\Phi(e_h), \Phi(e_t)\right),\ \Delta r(e_h, e_t),\ \cos\theta_{bearing}\right], \tag{12}$$

where d_K is the geodesic distance in $\mathbb{H}^d_K$ capturing physical proximity, $\Delta r = |r(\Phi(e_h)) - r(\Phi(e_t))|$ encodes the scale difference via radial depth gap, and $\cos\theta_{bearing}$ encodes the directional relationship. A lightweight MLP classifier f_θ then scores the plausibility of each candidate triple:

$$\text{score}(e_h, f^n_{spa}, e_t) = f_\theta(\mathbf{x}) \in [0, 1]. \tag{13}$$

Top-scored triples are retained and their tail entities form the frontier for next-hop expansion, following an adaptive search strategy with density-aware thresholding as in [18]. Specifically, we initialize with threshold $\tau_0 = 0.5$ and iteratively reduce by a decay factor $c = 0.1$ if the number of retrieved triples falls below a minimum acceptable count M, ensuring sufficient evidence coverage in sparse regions while preventing over-retrieval in dense regions.

3) Spatial Outward Einstein Midpoint Aggregation: After retrieval, the selected multi-source evidence typically spans multiple resolutions. To aggregate these into a unified representation without losing fine-scale information, we introduce the Spatial Outward Einstein Midpoint (Spatial OEM). The motivation stems from a known failure mode: naively averaging hyperbolic embeddings collapses representations toward the origin, destroying the hierarchical structure encoded in radial depth [20].

Given spatial observation hyperedge embeddings $\{\Phi(f_i)\}_{i=1}^n \subset \mathbb{H}^d_K$ with query-relevance weights w_i and resolution-aware radial weighting $\phi_{res}(f_i) = r(\Phi(f_i))^p$:

$$\mathbf{m}^{Spa\text{-}OEM}_{K,p} = \Pi_K\left(\frac{\sum_{i=1}^n w_i \cdot \phi_{res}(f_i) \cdot \lambda_i \cdot \Phi(f_i)}{\sum_{i=1}^n w_i \cdot \phi_{res}(f_i) \cdot \lambda_i}\right)$$

where $\lambda_i = \Phi(f_i)_0$ is the Lorentz factor and Π_K denotes reprojection onto $\mathbb{H}^d_K$, defined as $\Pi_K(\mathbf{v}) = \frac{\mathbf{v}}{\sqrt{K\langle\mathbf{v},\mathbf{v}\rangle_L}}$ for $\mathbf{v}$ with $\langle\mathbf{v},\mathbf{v}\rangle_L < 0$ and $v_0 > 0$.

Theorem 1 (Spatial OEM Outward Bias). *For $p \geq 1$, the Spatial OEM satisfies:*

$$r(\mathbf{m}^{Spa\text{-}OEM}_{K,p}) \geq r(\mathbf{m}^{Ein}_K)$$

where $\mathbf{m}^{Ein}_K$ is the standard Einstein midpoint ($p = 0$).

Proof. The OEM weights $\tilde{w}_i \propto w_i \cdot r(\Phi(f_i))^{p+1}$ concentrate more mass on high-radius points than the Einstein weights $w_i \cdot r(\Phi(f_i))$. By the Chebyshev sum inequality applied to the co-monotonic sequences $a_i = r(\Phi(f_i))^{p+1}$ and $b_i = r(\Phi(f_i))$, the pre-projection time component satisfies $\tilde{v}_0 \geq \bar{r}_w$ (weighted mean radius). Since reprojection Π_K preserves the ordering of time components, the result follows. □

The outward bias guarantees that high-resolution observations dominate the aggregated representation. This is essential for planetary science retrieval: when a user queries a specific geological feature, the aggregated evidence should preserve the fine-scale observational details rather than being smoothed into a coarse-resolution summary.

D. Physics-Informed Conflict Triage

We define the multi-source spatial evidence retrieved in a single query as observation-grounded homologous data (Definition 4). Although targeting the same query object, these data often provide inconsistent factual statements due to differences in instrument principles, observation geometry, and acquisition epochs. Unlike MultiRAG's Multi-level Confidence Computing (MCC), which assumes that inconsistency indicates unreliability and employs mutual information entropy to filter conflicting nodes, we adopt a fundamentally different paradigm: Physics-Informed Conflict Triage (PICT), which classifies conflicts by their physical origin and applies differentiated processing strategies.

1) Cross-Source Interaction Entropy: The first stage detects conflicts by measuring the information-theoretic interaction effect when two sources are jointly presented to the LLM. Existing entropy-based conflict detection methods, such as TruthfulRAG [17], compare retrieval-augmented entropy against parametric-only entropy ($\Delta H_p = H(P_{aug}) - H(P_{param})$). However, this formulation is inapplicable to our setting where all knowledge is external observational data rather than LLM parametric knowledge. We instead propose cross-source interaction entropy that measures the mutual interference between two observation sources:

$$\mathcal{H}_{inter}(p_i, p_j \mid q) = H\left(P(\text{ans} \mid q, p_i \oplus p_j)\right) - \frac{1}{2}\left[H\left(P(\text{ans} \mid q, p_i)\right) + H\left(\right.\right.$$

TABLE I
PHYSICS-INFORMED CONFLICT TRIAGE CATEGORIES

Category	Condition	Strategy
Noise ($\mathcal{C}^{noise}$)	Opaque, with significant source authority disparity	Filter low-authority source
Instrument-Inherent ($\mathcal{C}^{inst}$)	Explainable via $\Omega_i \neq \Omega_j$	Preserve with physical explanation
Scale-Dependent ($\mathcal{C}^{scale}$)	Explainable via $\ell^i_{res} \neq \ell^j_{res}$	Preserve with cross-scale linkage
Temporal-Evolution ($\mathcal{C}^{temp}$)	Explainable via $\mathcal{T}_i \neq \mathcal{T}_j$	Preserve with temporal ordering

where $H(\cdot)$ is the token-averaged entropy over top-k candidate tokens:

$$H\left(P(\text{ans} \mid \text{context})\right) = -\frac{1}{|l|}\sum_{t=1}^{|l|}\sum_{i=1}^{k} pr_i^{(t)} \log_2 pr_i^{(t)}$$

and $p_i \oplus p_j$ denotes the concatenation of both reasoning paths derived from sources $\mathcal{K}_i$ and $\mathcal{K}_j$ respectively. The interaction entropy admits a clear physical interpretation: positive values ($\mathcal{H}_{inter} > 0$, super-additive uncertainty) indicate that the two sources contradict each other, jointly creating more confusion than either alone; near-zero values indicate independence or consistency; negative values (sub-additive) indicate mutual complementarity where the sources reinforce each other.

Reasoning path pairs exhibiting interaction entropy exceeding a predefined threshold ϵ are classified as detected conflicts:

$$\mathcal{C}^{detected} = \{(\psi_i, \psi_j) \mid \mathcal{H}_{inter}(p_i, p_j \mid q) > \epsilon\}$$

2) Physics-Informed Conflict Classification: The second stage classifies each detected conflict by its physical origin. We introduce the central distinction of PICT:

Definition 7. Explainable conflict and opaque conflict. A pairwise conflict $(\psi_i, \psi_j) \in \mathcal{C}_{i,j}$ is *explainable* if there exists a physical bridging function $\mathcal{B}$ such that:

$$\mathcal{B}(\Omega_i, \Omega_j, \mathcal{M}_i, \mathcal{M}_j) \models \neg(\psi_i \perp \psi_j).$$

i.e., the apparent inconsistency is resolvable by accounting for observation constraint differences (Ω_i, Ω_j) and measurement model differences ($\mathcal{M}_i$, $\mathcal{M}_j$). Otherwise, the conflict is *opaque*.

Based on this distinction, we define four conflict categories, each with a differentiated processing strategy:

For each detected conflict, we construct a feature vector that fuses information-theoretic, physical, and neural signals:

$$\mathbf{z}_{conf} = \Big[\mathcal{H}_{inter},\ \|\Omega_i - \Omega_j\|,\ |\log(\ell^i_{res}/\ell^j_{res})|,\ \Delta\mathcal{T},\ \rho_{auth}(i,j),\ \mathbf{h}^{(l^*)}_{conf}\Big]$$

where $\|\Omega_i - \Omega_j\|$ is the observation geometry disparity, $|\log(\ell^i_{res}/\ell^j_{res})|$ is the resolution ratio in log-scale, $\Delta\mathcal{T}$ is the temporal separation, $\rho_{auth}(i,j)$ is the authority disparity between sources, and $\mathbf{h}^{(l^*)}_{conf}$ is the LLM hidden state at the conflict encoding layer. The inclusion of $\mathbf{h}^{(l^*)}_{conf}$ is motivated by the finding that knowledge conflict signals concentrate in mid-to-late layers of LLMs and are linearly separable with ¿ 93

A lightweight classifier maps the feature vector to conflict type:

$$\hat{c} = \arg\max_{c \in \{noise, inst, scale, temp\}} P_\theta(c \mid \mathbf{z}_{conf})$$

Proposition 2 (Conflict Type Separability). *The four conflict types are distinguished by orthogonal physical dimensions: $\|\Omega_i - \Omega_j\|$ separates instrument conflicts; $|\log(\ell^i_{res}/\ell^j_{res})|$ separates scale conflicts; $\Delta\mathcal{T}$ separates temporal conflicts; ρ_{auth} separates noise conflicts. Since these physical features are independent of and complementary to the hidden state features $\mathbf{h}^{(l^*)}_{conf}$ (which encode semantic inconsistency), the four conflict types are linearly separable in the augmented feature space $\mathbf{z}_{conf}$.*

3) Conflict-Aware Confidence Recalibration: Based on the classification result, we recalibrate the node confidence. This is the key departure from MultiRAG's MCC, which uniformly penalizes inconsistency:

$$C_{triage}(v) = \begin{cases} C_{base}(v) & \text{if } v \notin \mathcal{C}^{detected} \\ \alpha \cdot C_{base}(v) + (1-\alpha)\cdot\eta & \text{if } \hat{c} = noise \\ C_{base}(v) + \beta \cdot \mathcal{H}^{-1}_{inter} & \text{if } \hat{c} \in \{inst, scale\} \\ C_{base}(v) \cdot \gamma(|\Delta\mathcal{T}|) & \text{if } \hat{c} = temp \end{cases}$$

where $C_{base}(v)$ is the baseline confidence computed via semantic similarity (analogous to the node consistency score in [14]), $\eta < 0$ is a penalty term for noise conflicts, $\beta > 0$ is a boost coefficient for scientifically explainable conflicts, and $\gamma(|\Delta\mathcal{T}|)$ is a time-decay weighting function that prioritizes recent observations while preserving temporal evolution signals. Specifically, $\gamma(|\Delta\mathcal{T}|) = 1 + \beta_{temp} \cdot \exp(-|\Delta\mathcal{T}|/\tau_{decay})$, where $\beta_{temp} > 0$ ensures $\gamma > 1$ for temporal contrasts with scientific significance.

Theorem 2 (Anti-Over-Smoothing Guarantee). *Let $V_{sci} \subset V$ denote the set of nodes involved in explainable scientific conflicts ($\mathcal{C}^{inst} \cup \mathcal{C}^{scale} \cup \mathcal{C}^{temp}$). Under PICT with $\beta > 0$:*

$$C_{triage}(v) > C_{base}(v) \quad \forall v \in V_{sci}$$

Proof. For $v \in \mathcal{C}^{inst} \cup \mathcal{C}^{scale}$: $C_{triage}(v) = C_{base}(v) + \beta \cdot \mathcal{H}^{-1}_{inter}$. Since $\beta > 0$ and $\mathcal{H}_{inter} > \epsilon > 0$ (by the detection threshold in Eq. 16), $\beta \cdot \mathcal{H}^{-1}_{inter} > 0$, thus $C_{triage}(v) > C_{base}(v)$. For $v \in \mathcal{C}^{temp}$: $\gamma(|\Delta\mathcal{T}|) > 1$ by construction (since $\beta_{temp} > 0$ and $\exp(\cdot) > 0$), thus $C_{triage}(v) = C_{base}(v) \cdot \gamma(|\Delta\mathcal{T}|) > C_{base}(v)$. □

This theorem provides a formal guarantee that scientifically valuable conflict nodes can never be suppressed below their baseline confidence by the triage mechanism, directly addressing the over-smoothing problem.

E. AreoRAG Prompting

We propose the AreoRAG Prompting (ARP) algorithm for multi-source planetary spatial data retrieval. The complete procedure is presented in Algorithm 1.

Algorithm 1 AreoRAG Prompting (ARP)

Require: Query q
Ensure: Generated Answer
$\mathcal{E}_q, \mathcal{R}_q, \mathcal{P}_{foot}, \mathcal{T}_{win} \leftarrow$ Spatial Intent Extraction(q)
$D_q \leftarrow$ Multi-source Spatial Adapter Parsing(D) {Eq. 4–5}
$\mathcal{G}_{hyp} \leftarrow$ HySH Construction(D_q) {Eq. 6–8}
$\mathcal{T}_q \leftarrow$ Spatiotemporal Retrieval$(\mathcal{G}_{hyp}, \mathcal{E}_q)$ {Eq. 9–12}
$\mathbf{m}_{agg} \leftarrow$ Spatial OEM Aggregation$(\mathcal{T}_q)$ {Eq. 13}
$\mathcal{C}^{detected} \leftarrow$ Cross-Source Interaction Entropy$(\mathcal{T}_q, q)$ {Eq. 14–16}
for $(\psi_i, \psi_j) \in \mathcal{C}^{detected}$ **do**
  $\hat{c} \leftarrow$ Conflict Classification$(\mathbf{z}_{conf})$ {Eq. 18–19}
  $C_{triage}(v) \leftarrow$ Confidence Recalibration$(v, \hat{c})$ {Eq. 20}
end for
Context $\leftarrow$ Differential Context Construction$(q, \mathcal{T}_q, \hat{c})$
Answer $\leftarrow$ LLM$(q \oplus$ Context $\oplus$ Provenance$)$
return Answer

Given a user query q, the LLM is first employed to extract entities, spatial constraints ($\mathcal{P}_{foot}$, region), and temporal constraints ($\mathcal{T}_{win}$, L_s range), generating corresponding logical and spatial relationships. The observation data then undergoes multi-source spatial adapter parsing to derive normalized datasets (Eq. 4), followed by constructing a Hyperbolic Spatial Hypergraph via scale-aware Lorentz embedding and cross-reference-frame alignment (Eq. 6-8).

Subsequently, spatiotemporal retrieval is performed using hyperbolic spatial encoding and MLP-based plausibility scoring (Eq. 10-12), with Spatial OEM aggregation (Eq. 13) to produce a unified cross-resolution representation. The cross-source interaction entropy mechanism (Eq. 14-16) then detects inter-source conflicts, after which each detected conflict is classified via the physics-informed feature vector (Eq. 18-19) and the node confidence is recalibrated accordingly (Eq. 20).

The final step constructs a differential context based on the triage result. For noise conflicts, the low-authority source is filtered, compatible with conventional conflict elimination. For instrument-inherent and scale-dependent conflicts, both sources are preserved with a physical bridging explanation $\mathcal{B}(\Omega_i, \Omega_j)$ appended to the context, enabling the LLM to reason about the physical origin of the disagreement. For temporal-evolution conflicts, a temporal ordering is constructed, allowing the LLM to trace the evolution of observations over time. All preserved evidence carries provenance metadata (DataID, source institution, instrument identity, observation timestamp in L_s) to ensure scientific traceability, analogous to the citation anchors in Perplexity-style retrieval systems.

It should be noted that the ARP algorithm constructs the HySH offline as a preprocessing step, while the PICT module operates online during each query. The HySH construction time is dominated by the LLM-based entity extraction (comparable to MultiRAG's MLG construction), while the online PICT overhead consists primarily of $|\mathcal{C}^{detected}|$ forward passes through the lightweight conflict classifier (Eq. 19), which is negligible compared to the LLM generation cost.

IV. Experiments

This section conducts experiments and performance analysis on the Hyperbolic Spatial Hypergraph (HySH) construction and the Physics-Informed Conflict Triage (PICT) modules. Baseline methods are compared with SOTA multi-source retrieval, graph-based RAG, and conflict-resolution methods. Extensive experiments are conducted to assess the robustness and efficiency of AreoRAG, which aims to answer the following questions.

- **Q1**: How does the overall retrieval and QA performance of AreoRAG compare with existing multi-source RAG and graph-based RAG methods on planetary spatial data?
- **Q2**: What are the respective impacts of spatial sparsity and inter-source conflict intensity on retrieval quality?
- **Q3**: How effective are the two core modules (HySH and PICT) of AreoRAG individually?
- **Q4**: Can PICT correctly preserve scientifically valuable conflicts while filtering noise, and how does this compare with conventional conflict-elimination approaches?
- **Q5**: What are the time costs of the various modules in AreoRAG?

A. Experimental Settings

a) Datasets: To validate the effectiveness of AreoRAG in planetary multi-source spatial data retrieval, we construct three datasets from real Mars exploration archives and further evaluate on two general multi-hop QA benchmarks. The planetary datasets are summarized in Table I.

(1) **MarsRegion-QA**: A multi-source spatial QA dataset constructed from the Mars Orbital Data Explorer (ODE) archives. We select five scientifically significant regions on Mars — Jezero Crater, Gale Crater, Utopia Planitia (Zhurong landing site), Valles Marineris, and Olympus Mons — and aggregate observations from HiRISE (0.3 m), CTX (6 m), CRISM (18 m), MOLA (460 m), and Zhurong/Curiosity rover in-situ measurements. Each query targets cross-source spatial reasoning (e.g., "What mineral signatures have been detected in the clay-bearing unit at the western delta of Jezero Crater, and do orbital and in-situ observations agree?"). We construct 200 queries with expert-annotated ground truth answers and conflict labels.

(2) **MarsConflict-50**: A curated subset of 50 observation pairs exhibiting known scientific conflicts documented in the planetary science literature (e.g., orbital detection of hydrated minerals vs. inconclusive in-situ results). Each pair is annotated with conflict type (instrument-inherent, scale-dependent, temporal-evolution, or noise) by domain experts. This dataset serves as the primary benchmark for evaluating PICT's conflict classification accuracy.

(3) **MarsTemporal-QA**: A temporal reasoning dataset comprising 150 queries about surface changes observed across different Mars Years (MY), such as recurring slope lineae (RSL) activity, dust storm impacts, and seasonal frost patterns. Each query requires integrating observations spanning L_s ranges to assess temporal evolution.

Additionally, to validate generalization on established benchmarks, we evaluate on HotpotQA [38] and 2WikiMul-

TABLE II
STATISTICS OF THE PLANETARY DATASETS

Dataset	Data Source	Sources	Entities	Hyperedges	Queries
MarsRegion-QA	HiRISE (Orbital)	1	12,847	8,213	200
	CTX (Orbital)	1	28,563	15,471	
	CRISM (Orbital)	1	6,329	4,182	
	MOLA (Orbital)	1	45,210	22,605	
	Rover In-situ	2	3,876	2,541	
MarsConflict-50	Mixed (all above)	6	1,247	683	50
MarsTemporalQA	Mixed (all above)	6	8,934	5,127	150

tiHopQA [39], using the same 300-question subsamples as MultiRAG [14] for fair comparison.

It is noteworthy that MarsRegion-QA exhibits high spatial density (multiple overlapping observations per region) but significant cross-resolution heterogeneity, while MarsConflict-50 is specifically designed to stress-test conflict handling with a high proportion of scientifically valuable disagreements (72% of conflicts are non-noise).

b) Evaluation Metrics: We adopt multiple metrics to comprehensively evaluate retrieval quality, answer accuracy, and conflict handling:

- **F1 score**: The harmonic mean of precision and recall, assessing overall retrieval and answer quality:

$$F1 = 2 \times \frac{P \times R}{P + R}$$

- **Recall@K**: Recall at rank K, measuring the proportion of relevant documents retrieved within the top-K results.

- **Conflict Preservation Rate (CPR)**: The proportion of scientifically valuable conflicts (annotated as instrument-inherent, scale-dependent, or temporal-evolution) that are correctly preserved rather than filtered:

$$CPR = \frac{|\mathcal{C}^{sci}_{preserved}|}{|\mathcal{C}^{sci}_{total}|}$$

- **Noise Rejection Rate (NRR)**: The proportion of noise conflicts that are correctly filtered:

$$NRR = \frac{|\mathcal{C}^{noise}_{filtered}|}{|\mathcal{C}^{noise}_{total}|}$$

- **Conflict Classification Accuracy (CCA)**: Four-class classification accuracy over the conflict types on MarsConflict-50.

- **Query Time (QT)** and **Preprocessing Time (PT)**: Measured in seconds, assessing online and offline efficiency.

c) Hyper-parameter Settings: All methods were implemented in Python 3.10 and CUDA 12.1 environment. The base LLM is Llama3-8B-Instruct for all methods except where noted. For HySH construction, the hyperbolic curvature is set to $K = -1.0$, the embedding dimension $d = 64$, and the resolution power parameter $p = 2$ for Spatial OEM. For PICT, the interaction entropy threshold is $\epsilon = 0.3$, the noise penalty $\eta = -0.5$, the scientific boost coefficient $\beta = 0.2$, the temporal decay constant $\tau_{decay} = 180$ (in L_s degrees, approximately one Mars season), and the authority weight $\alpha = 0.5$. The MLP conflict classifier uses a two-layer architecture ($256 \to 128 \to 4$) with ReLU activation, trained on MarsConflict-50 with 5-fold cross-validation. The plausibility scoring MLP f_θ for retrieval follows the architecture in [18] with adaptive threshold $\tau_0 = 0.5$ and decay factor $c = 0.1$. All experiments were conducted on a device equipped with an NVIDIA A100 (80 GB) GPU and 256 GB of memory.

d) Baseline Models: To demonstrate the superiority of AreoRAG, we compare with the following categories of methods:

General RAG Methods:

1) **Standard RAG** [6]: Conventional retrieval-augmented generation with dense vector retrieval.

2) **IRCoT** [44]: Iterative retrieval with chain-of-thought reasoning refinement.

3) **RQ-RAG** [47]: Retrieval with optimized query decomposition for complex queries.

Graph-based RAG Methods:

4) **MultiRAG** [14]: Multi-source line graph with multi-level confidence computing (the primary comparison target).

5) **HyperGraphRAG** [25]: Hypergraph-based RAG with n-ary relational facts retrieval.

6) **HyperRAG** [18]: MLP-based retrieval over n-ary hypergraphs with adaptive search.

Conflict-Resolution Methods:

7) **TruthfulRAG** [17]: Knowledge graph-based conflict resolution via entropy-based filtering.

8) **MetaRAG** [9]: Metacognitive strategies for hallucination mitigation in retrieval.

e) Dataset Preprocessing: For the planetary datasets, we parse PDS4 labels and CNSA metadata through the multi-source spatial adapters (Section III-B) to extract spatial footprints, temporal windows, and instrument parameters. All observations are projected to the Mars IAU 2000 areocentric coordinate system. Temporal references are unified to Solar Longitude L_s using SPICE kernels. For the general QA benchmarks, we follow the same preprocessing pipeline as MultiRAG [14] to ensure fair comparison.

B. Overall Retrieval and QA Performance (Q1)

To validate the effectiveness of AreoRAG, we assess it using F1 scores and query times across the planetary datasets and the two general multi-hop QA benchmarks. Table II summarizes the performance comparison.

Table II demonstrates that AreoRAG outperforms all comparative methods across both planetary and general QA

TABLE III
COMPARISON WITH BASELINE METHODS ON PLANETARY AND GENERAL QA DATASETS

Method	MarsRegion-QA		MarsTemporal-QA		HotpotQA		2WikiMultiHopQA	
	F1/%	Recall@5	F1/%	Recall@5	F1/%	Recall@5	F1/%	Recall@5
Standard RAG	28.4	31.2	25.7	28.3	34.1	33.5	25.6	26.2
IRCoT	35.6	38.9	32.1	35.4	41.6	41.2	42.3	40.9
RQ-RAG	37.2	40.5	34.8	37.6	51.6	49.3	45.3	44.6
MultiRAG	42.3	46.8	38.5	42.1	59.3	62.7	55.7	61.2
HyperGraphRAG	44.1	48.3	40.2	43.7	51.0	42.7	42.5	30.2
HyperRAG	46.5	50.7	41.8	45.2	42.5	43.7	34.0	34.1
TruthfulRAG	40.8	44.6	37.9	41.3	60.2	—	55.4	—
MetaRAG	41.5	45.2	39.1	42.8	51.1	49.9	50.7	52.2
AreoRAG	**55.8**	**61.3**	**52.4**	**57.6**	**61.7**	**64.2**	**57.3**	**62.8**

datasets. On MarsRegion-QA, AreoRAG achieves an F1 score of 55.8%, representing a 13.5% absolute improvement over MultiRAG (42.3

On MarsTemporal-QA, which demands temporal reasoning across observation epochs, AreoRAG achieves 52.4% F1, outperforming all baselines by at least 10.6%. This improvement is attributed to PICT's temporal-evolution conflict handling (the $\gamma(|\Delta\mathcal{T}|)$ weighting in Eq. 20), which preserves temporal change signals rather than filtering them as inconsistencies.

On the general benchmarks (HotpotQA and 2WikiMultiHopQA), AreoRAG maintains competitive performance (61.7% and 57.3% F1), demonstrating that the framework generalizes beyond planetary science. The modest improvements over MultiRAG on these benchmarks (2.4% and 1.6%) are expected, as these datasets do not exhibit the spatial and physical conflict characteristics that AreoRAG is specifically designed to address.

Notably, HyperRAG and HyperGraphRAG perform well on planetary datasets (46.5% and 44.1% F1 on MarsRegion-QA) but underperform on general benchmarks. This is because their n-ary hypergraph structure naturally accommodates the multi-entity spatial observations in planetary data, yet they lack the conflict triage mechanism needed to handle inter-source disagreements correctly.

C. Robustness Under Spatial Sparsity and Conflict Intensity (Q2)

AreoRAG demonstrates strong robustness under varying spatial sparsity and conflict intensity. We conduct experiments from two perspectives.

1) Spatial Sparsity: We applied 30%, 50%, and 70% random hyperedge masking to MarsRegion-QA, progressively removing spatial connections while ensuring query answers remain retrievable.

As shown in Fig. 5(a-b), after applying 30%, 50%, and 70% hyperedge masking, AreoRAG's F1 score on MarsRegion-QA decreased from 55.8% to 52.1%, 49.3%, and 45.6% respectively. In contrast, MultiRAG's F1 dropped more sharply from 42.3% to 37.8%, 32.5%, and 26.1%. HyperRAG shows moderate degradation (46.5% to 42.7%, 38.9%, 33.4%). The superior robustness of AreoRAG under sparsity is attributed to two factors: (i) hyperbolic embedding preserves proximity information even when explicit graph edges are removed, as geodesic distance in $\mathbb{H}^d_K$ encodes spatial proximity independently of graph connectivity; and (ii) the Spatial OEM aggregation maintains representational quality by amplifying high-resolution signals that survive masking.

2) Conflict Intensity: We injected 30%, 50%, and 70% synthetic conflict triples into MarsRegion-QA by duplicating existing observation records and perturbing their factual content (e.g., randomizing mineral identifications or altering coordinate data), simulating scenarios of increasing inter-source noise.

As shown in Fig. 5(c-d), AreoRAG's F1 score decreased only moderately from 55.8% to 54.2%, 52.8%, and 50.1% under 30%, 50%, and 70% conflict injection respectively. MultiRAG exhibited steeper degradation (42.3% to 40.1%, 36.4%, 30.7%), and TruthfulRAG showed similar sensitivity (40.8% to 38.2%, 34.6%, 29.3%). The resilience of AreoRAG is directly attributable to PICT's ability to classify injected noise conflicts as $\mathcal{C}^{noise}$ and filter them while preserving genuine scientific disagreements. In contrast, MultiRAG's MCC module and TruthfulRAG's entropy-based filtering indiscriminately penalize all inconsistencies, including the original valid observations that become "outvoted" by injected noise.

D. Ablation Study (Q3)

To evaluate the individual contributions of HySH and PICT, we conduct systematic ablation experiments. Table III reports results on MarsRegion-QA and MarsTemporal-QA.

a) HySH Module Analysis: The HySH module achieves significant improvements in both accuracy and efficiency. Replacing HySH with MultiRAG's MLG (w/o HySH) causes F1 drops of 11.2% on MarsRegion-QA and 12.3% on MarsTemporal-QA, while query time increases by 8.4× (3.42s to 28.7s) due to the edge explosion problem in pairwise spatial encoding. This validates the $O(k)$ vs. $O(k^2)$ complexity advantage of hyperedges.

Within HySH, the hyperbolic embedding contributes 6.6% F1 improvement over Euclidean hypergraph (49.2% vs. 55.8%), confirming that the negative-curvature geometry is essential for faithfully representing the hierarchical scale structure. The Spatial OEM contributes an additional 4.5% F1 over standard Einstein midpoint aggregation (51.3% vs. 55.8%), validating the outward bias property (Theorem 1) in preventing hierarchical collapse during cross-resolution fusion.

TABLE IV
ABLATION EXPERIMENTS OF HYSH AND PICT MODULES

Configuration	MarsRegion-QA			MarsTemporal-QA		
	F1/%	QT/s	PT/s	F1/%	QT/s	PT/s
AreoRAG (Full)	55.8	3.42	86.5	52.4	4.17	72.3
w/o HySH (use MLG)	44.6	28.7	15.2	40.1	35.4	12.8
w/o Hyperbolic (Euclidean hypergraph)	49.2	4.85	51.3	45.6	5.72	43.7
w/o Spatial OEM (standard Einstein)	51.3	3.38	86.5	47.8	4.12	72.3
w/o PICT (use MCC)	45.9	3.15	86.5	39.7	3.89	72.3
w/o Conflict Classification (uniform filter)	48.1	3.28	86.5	42.3	4.01	72.3
w/o Interaction Entropy (use ΔH_p)	50.4	3.51	86.5	46.2	4.25	72.3
w/o Both (Standard RAG)	28.4	1.23	—	25.7	1.56	—

b) PICT Module Analysis: Replacing PICT with MultiRAG's MCC (w/o PICT) causes F1 drops of 9.9% on MarsRegion-QA and 12.7% on MarsTemporal-QA. The larger drop on MarsTemporal-QA is expected, as this dataset contains abundant temporal-evolution conflicts that MCC would filter as inconsistencies.

The ablation further reveals the contribution of each PICT component. Removing conflict classification (using uniform filtering instead of four-category triage) costs 7.7% F1 on MarsRegion-QA. Replacing cross-source interaction entropy with TruthfulRAG's ΔH_p metric costs 5.4% F1, confirming that the cross-source formulation (Eq. 14) is more appropriate for the all-external-knowledge setting of planetary observations.

c) Module Interaction: Notably, the sum of individual module contributions (HySH: 11.2% + PICT: 9.9% = 21.1%) exceeds the gap between the full model and Standard RAG (55.8% - 28.4% = 27.4%), but the actual synergy is evident in the coupling points. HySH's radial depth difference Δr directly improves PICT's scale-conflict classification; PICT's triage feedback improves HySH's retrieval priority. Disabling either module degrades the other's performance more than isolated analysis suggests.

E. Conflict Preservation Evaluation (Q4)

A defining capability of AreoRAG is the ability to preserve scientifically valuable conflicts rather than suppressing them. We evaluate this on MarsConflict-50, which contains expert-annotated conflict types.

Standard RAG preserves all information indiscriminately (CPR=100%) because it has no conflict handling mechanism, resulting in noise contamination and low F1. "—" indicates the method does not perform explicit conflict classification.

Table IV reveals the fundamental difference between AreoRAG and existing methods. MultiRAG achieves a high Noise Rejection Rate (85.7%) but at the cost of a catastrophically low Conflict Preservation Rate (8.3%) — it filters 91.7% of scientifically valuable conflicts as "unreliable data." TruthfulRAG and MetaRAG show similar behavior (CPR of 13.9% and 11.1%), confirming that existing conflict-resolution methods systematically destroy scientific anomaly signals.

TABLE V
CONFLICT HANDLING PERFORMANCE ON MARSCONFLICT-50

Method	CCA/%	CPR/%	NRR/%	F1/%
Standard RAG	—	100.0*	0.0	26.3
MultiRAG (MCC)	—	8.3	85.7	35.2
TruthfulRAG	—	13.9	78.6	37.8
MetaRAG	—	11.1	82.1	36.5
AreoRAG (PICT)	**84.0**	**91.7**	**85.7**	**53.1**

In contrast, AreoRAG achieves a CPR of 91.7% while maintaining the same NRR (85.7%) as MultiRAG, demonstrating that PICT successfully decouples noise filtering from scientific conflict preservation. The Conflict Classification Accuracy of 84.0% on the four-category task validates the separability claim in Proposition 2. Error analysis reveals that the primary source of misclassification is between instrument-inherent and scale-dependent conflicts (12.3% confusion rate), which is expected as both involve observation geometry differences. Noise vs. scientific conflict misclassification is rare (3.7%), confirming the robustness of the explainable/opaque distinction (Definition 7).

Furthermore, the F1 score improvement (53.1% vs. 35.2% for MultiRAG) demonstrates that preserving scientific conflicts directly benefits answer quality: the LLM can generate more comprehensive and scientifically faithful answers when provided with both agreeing and legitimately disagreeing evidence, accompanied by physical bridging explanations.

F. Efficiency Analysis (Q5)

AreoRAG's query time (3.42s on MarsRegion-QA) is competitive with HyperRAG (2.95s) and substantially faster than MultiRAG (4.87s) and TruthfulRAG (5.62s). The faster online query is attributable to the $O(k)$ hyperedge traversal complexity and the lightweight MLP-based plausibility scoring, which avoids the expensive mutual information entropy computation required by MultiRAG's MCC at query time.

The preprocessing time (86.5s) is higher than MultiRAG (15.2s) due to the hyperbolic embedding computation (Eq.

TABLE VI
TIME COST ANALYSIS ACROSS MODULES

Method	MarsRegion-QA		MarsTemporal-QA	
	QT/s	PT/s	QT/s	PT/s
Standard RAG	1.23	—	1.56	—
MultiRAG	4.87	15.2	6.13	12.8
HyperRAG	2.95	142.7	3.41	118.5
TruthfulRAG	5.62	18.7	6.85	15.4
AreoRAG	**3.42**	**86.5**	**4.17**	**72.3**

6-8), but lower than HyperRAG (142.7s) because we do not require the full contrastive training pipeline. Importantly, HySH construction is a one-time offline cost amortized across all queries. The PICT module adds minimal online overhead: the conflict classifier (Eq. 19) requires <0.1s per detected conflict pair, and the interaction entropy computation (Eq. 14) adds approximately 0.8s per query through parallel LLM forward passes.

G. Case Study

AreoRAG's effectiveness in multi-source planetary data integration is demonstrated through a real-world query about the Jezero Crater western delta. The query and system response are detailed in Table VI.

This case study exemplifies AreoRAG's core advantage: while MultiRAG filters the in-situ observation as "unreliable" due to its inconsistency with orbital data, AreoRAG recognizes this as a scale-dependent conflict, preserves both observations, and generates a scientifically meaningful explanation (spatial mixing effect). The answer includes provenance metadata (DataIDs) for scientific traceability, and proactively recommends follow-up data to resolve the ambiguity — a capability enabled by the PICT module's conflict-aware context construction.

H. Limitations

We acknowledge several limitations inherent in the current framework:

1) **Dataset scale**: The planetary datasets are constructed from publicly available archives and may not cover the full diversity of Mars exploration scenarios. Larger-scale evaluation with comprehensive PDS holdings is planned as future work.

2) **Conflict classification coverage**: The four-category conflict taxonomy, while covering the most common planetary science scenarios, may not capture all possible conflict origins (e.g., processing artifact conflicts, calibration drift). Extending the taxonomy is a natural direction.

3) **LLM dependency**: The cross-source interaction entropy computation (Eq. 14) and conflict classification (Eq. 18) both rely on LLM forward passes, introducing potential biases from the base model's parametric knowledge about planetary science. Fine-tuning on domain-specific corpora may mitigate this issue.

4) **Generalization to other planetary bodies**: While designed for Mars, the framework's principles (hyperbolic scale hierarchy, physics-informed conflict triage) are applicable to other planetary bodies (Moon, Venus, icy moons). Validation on non-Mars datasets remains future work.

I. Related Work

J. Graph-Structured Retrieval Augmented Generation

Graph-based methods have become a central paradigm for enhancing the reasoning capabilities and factual grounding of Retrieval Augmented Generation (RAG) systems. Early approaches leveraged curated Knowledge Graphs (KGs) such as Wikidata and Freebase to provide structured triples or reasoning chains for LLM-based question answering [22], [27], [40]. More recently, methods that dynamically construct task-specific graphs from raw corpora have gained prominence. HippoRAG [23] draws inspiration from neurobiology to construct offline memory graphs with a neural indexing mechanism, achieving significant retrieval latency reduction. ToG 2.0 [25] introduces a graph-context co-retrieval framework that dynamically balances structured and unstructured evidence, resulting in substantial hallucination rate reduction compared to unimodal approaches. Graph-CoT [48] leverages Graph Neural Networks to establish bidirectional connections between KGs and the latent space of LLMs, reducing factual inconsistencies on KGQA benchmarks. SubGraphRAG [19] proposes a lightweight MLP-based approach that retrieves query-relevant subgraphs and encodes structural proximity through directional distance encoding, achieving state-of-the-art performance with low latency.

A critical limitation of the above methods is their reliance on binary relational facts (entity-relation-entity triples), which suffer from semantic fragmentation and path explosion when representing complex multi-entity interactions [18]. To address this, hypergraph-based RAG methods have emerged. HyperGraphRAG [25b] advances the field by natively encoding n-ary relational facts as hyperedges, outperforming conventional KG-based RAGs through shallower yet more expressive reasoning chains. HyperRAG [18] further introduces a trainable MLP-based retriever (HyperRetriever) that fuses structural and semantic signals for adaptive n-ary chain construction, achieving the highest answer accuracy on WikiTopics benchmarks. OG-RAG [34b] grounds hyperedge construction in domain-specific ontologies for more interpretable evidence aggregation, though its dependence on high-quality ontologies constrains scalability.

For multi-source scenarios, MultiRAG [14] proposes multi-source line graphs (MLG) to aggregate cross-domain knowledge and multi-level confidence computing (MCC) to filter unreliable nodes, achieving over 10% F1 improvement on sparse datasets. FusionQuery [34] enhances cross-domain retrieval precision through heterogeneous graph integration with dynamic credibility evaluation. KAG [26] provides a unified representation framework for multi-source KGs through the OpenSPG platform.

Despite this progress, all existing graph-based RAG methods — whether binary, hypergraph, or multi-source line graph — construct their topology based on discrete text entities and explicit semantic associations. None addresses the scenario where data sources are inherently embedded in continuous

physical space and where inter-entity relevance is governed by spatial proximity rather than textual co-occurrence. AreoRAG bridges this gap by introducing spatial observation hyperedges embedded in hyperbolic space, enabling faithful representation of continuous spatiotemporal topology within a graph-based retrieval framework.

K. Hyperbolic Representation Learning for Retrieval

Hyperbolic geometry has attracted increasing attention in representation learning due to its capacity to embed hierarchical, tree-like structures with low distortion [52]-[54]. Unlike Euclidean space, where volume grows polynomially with radius, hyperbolic space exhibits exponential volume growth, naturally accommodating the branching structure of taxonomies, ontologies, and scale hierarchies. Foundational work by Nickel and Kiela [52] demonstrated that Poincaré embeddings of WordNet hierarchies achieve superior link prediction with substantially fewer dimensions than Euclidean counterparts. Subsequent work extended hyperbolic representations to knowledge graph embedding [53], [55], molecular generation [56], and recommendation systems [57].

In the context of text retrieval, hyperbolic geometry has recently shown strong promise. HypRAG [20] introduces hyperbolic dense retrieval for RAG, developing two model variants in the Lorentz model: a fully hyperbolic transformer (HyTE-FH) and a hybrid architecture (HyTE-H). A key contribution is the Outward Einstein Midpoint (OEM), a geometry-aware pooling operator that provably preserves hierarchical structure during sequence aggregation, overcoming the radial contraction failure of naive Euclidean averaging. HypRAG achieves up to 29% gains over Euclidean baselines in context relevance on RAGBench, and demonstrates that hyperbolic representations encode document specificity through norm-based separation — with over 20% radial increase from general to specific concepts. HyperbolicRAG [58] projects embeddings into the Poincaré ball to encode hierarchical depth within a static knowledge graph, using dual-space retrieval that fuses Euclidean and hyperbolic rankings. HELM [59] introduces a family of hyperbolic language models that operate entirely in hyperbolic space for text generation, though not specifically targeting retrieval.

These works establish the viability of hyperbolic geometry for hierarchical text retrieval, but they exclusively address the semantic hierarchy of natural language documents (broad topics $\rightarrow$ specific entities). No existing work has applied hyperbolic geometry to represent the physical scale hierarchy of scientific observations, where the hierarchy arises not from semantic abstraction but from spatial resolution (coarse global survey $\rightarrow$ fine local imaging). AreoRAG introduces the scale-curvature correspondence principle (Proposition 1), which establishes that the resolution hierarchy of planetary remote sensing data is intrinsically hyperbolic, and couples spatial resolution with radial depth in the Lorentz model. Furthermore, we extend the OEM pooling operator with resolution-aware radial weighting (Spatial OEM, Eq. 13), ensuring that cross-resolution aggregation preserves fine-scale observational details rather than collapsing them into coarse-resolution summaries.

L. Knowledge Conflict Detection and Resolution in RAG

Knowledge conflicts — situations where different information sources provide contradictory factual statements — pose a fundamental challenge to RAG systems [60]-[62]. Research on conflict handling can be broadly categorized into impact analysis and resolution strategies.

Impact analysis. Longpre et al. [60] first exposed entity-based knowledge conflicts in question answering, revealing that LLMs tend to rely on parametric memory when retrieved passages contain contradictory information. Xie et al. [61] found that LLMs are receptive to single external evidence but exhibit strong confirmation bias when presented with both supporting and conflicting information. Tan et al. [63] revealed a systematic bias toward self-generated contexts over retrieved ones, attributing this to higher query-context similarity of self-generated content. More recently, Tang et al. [21] formalized knowledge conflict in multimodal long-chain reasoning, distinguishing between input-level objective conflict and process-level effective conflict. Through probing internal representations, they revealed four key findings: (I) different conflict types are encoded as linearly separable features (¿93% AUC with linear probes); (II) conflict signals concentrate in mid-to-late layers (depth localization); (III) aggregating token-level signals along trajectories robustly recovers input-level conflict types (hierarchical consistency); and (IV) reinforcing the model's implicit source preference is far easier than reversing it (directional asymmetry). These mechanistic insights provide the theoretical foundation for PICT's conflict classification approach.

Resolution strategies. Existing resolution methods operate at the token level or semantic level [64]-[67]. Token-level methods such as CD^2 [64] manipulate attention weights to suppress parametric knowledge when conflicts are detected. ASTUTE RAG [65] uses gradient-based attribution to identify and mask conflicting tokens during inference. Semantic-level methods include CK-PLUG [66], which develops adapter-based architectures for dynamic knowledge weighting, and FaithfulRAG [67], which externalizes LLMs' parametric knowledge and aligns it with retrieved context. TruthfulRAG [17] advances to factual-level resolution by constructing knowledge graphs from retrieved content, performing query-based graph retrieval, and applying entropy-based filtering to locate conflicting elements — specifically comparing retrieval-augmented entropy against parametric-only entropy (ΔH_p) to identify corrective knowledge paths. MetaRAG [9] employs metacognitive strategies for hallucination mitigation through self-reflection mechanisms.

A critical and unexamined assumption shared by all existing conflict-resolution methods is that inter-source inconsistency is inherently undesirable and should be eliminated. This assumption holds in domains where authoritative ground truth exists (e.g., financial records, encyclopedic facts). However, in scientific observation scenarios — particularly deep-space exploration — the absence of absolute ground truth means that inter-source disagreements may represent legitimate multi-dimensional observations of the same phenomenon rather than errors. AreoRAG introduces a fundamentally different

paradigm: Physics-Informed Conflict Triage (PICT), which classifies conflicts by their physical origin and applies differentiated processing. By replacing TruthfulRAG's parametric-vs-augmented entropy (ΔH_p) with cross-source interaction entropy ($\mathcal{H}_{inter}$, Eq. 14) and incorporating physical observation parameters alongside LLM hidden-state features for four-category conflict classification (Eq. 18-19), PICT provably preserves scientifically valuable disagreements (Theorem 2) while maintaining noise-filtering capability.

M. Intelligent Retrieval for Planetary Remote Sensing Data

Planetary remote sensing archives have grown to petabyte scale through missions such as Mars Reconnaissance Orbiter, Mars Express, Tianwen-1, Mars Science Laboratory, and Mars 2020 [1]-[4]. The primary access infrastructure — NASA's Planetary Data System (PDS) [68] and its Mars Orbital Data Explorer (ODE) [69] — provides metadata-driven search through spatial bounding box queries, temporal range filters, and instrument/product-type selectors. Similarly, CNSA's Lunar and Planetary Data Release System offers keyword-based retrieval for Chinese mission data [70]. The USGS Astrogeology Science Center maintains derived data products (DTMs, mosaics) with catalog-level metadata search [71].

However, these systems operate at the level of metadata keyword matching and do not support semantic understanding of query intent, cross-source reasoning, or natural language interaction. A scientist seeking "HiRISE images showing dust devil tracks near the equator" must manually translate this into a series of coordinate-bounded, instrument-filtered queries and visually inspect each returned product — a process that is both labor-intensive and prone to missing relevant observations cataloged under different terminology.

In the broader geospatial domain, the integration of AI with remote sensing data retrieval has gained momentum. GeoAI methods [72], [73] combine geographic information science with deep learning for tasks such as scene classification, object detection, and change detection. Recent work has explored the use of LLMs for geospatial reasoning [74], [75], including natural language interfaces for GIS queries and the interpretation of satellite imagery through vision-language models. Foundation models for remote sensing, such as those pre-trained on large-scale Earth observation data, have demonstrated the potential for cross-modal understanding [76], [77]. However, these efforts remain focused on Earth observation data and do not address the unique challenges of planetary science: the multi-platform observation geometry, the absence of ground truth for conflict adjudication, and the need for cross-resolution reasoning across vastly different spatial scales.

To the best of our knowledge, AreoRAG is the first framework that brings RAG capabilities to planetary remote sensing data retrieval. By constructing a spatially-grounded knowledge hypergraph with physics-informed conflict handling, AreoRAG transforms the planetary data retrieval paradigm from metadata keyword matching to semantic spatial reasoning, enabling natural language queries that involve spatial proximity, temporal evolution, cross-source correlation, and scientifically informed conflict interpretation.

V. Conclusion

In this work, we introduce AreoRAG, a framework designed for multi-source planetary spatial data retrieval augmented generation. To address the structural bottleneck of discrete representation failure for continuous spatiotemporal topology and the epistemological conflict between scientific observational divergence and traditional de-falsification mechanisms, we propose two key innovations: Hyperbolic Spatial Hypergraph construction and Physics-Informed Conflict Triage.

The introduction of HySH employs n-ary spatial observation hyperedges embedded in hyperbolic space via the Lorentz model, reducing edge complexity from $O(k^2)$ to $O(k)$ while faithfully preserving the hierarchical scale structure of planetary observations through the scale-curvature correspondence principle. The Spatial Outward Einstein Midpoint aggregation operator further ensures that cross-resolution evidence fusion retains fine-scale observational details with a formal outward bias guarantee. Meanwhile, the PICT module fundamentally redefines the role of inter-source conflict in RAG systems — shifting from uniform conflict elimination to physics-informed conflict triage that classifies disagreements by their physical origin and applies differentiated confidence recalibration. The Anti-Over-Smoothing Guarantee (Theorem 2) ensures that scientifically valuable observational divergences are provably preserved rather than suppressed.

Extensive experiments on multi-source planetary observation datasets and general multi-hop QA benchmarks demonstrate that AreoRAG significantly outperforms existing methods in retrieval fidelity, answer accuracy, and scientific faithfulness. In particular, AreoRAG achieves a Conflict Preservation Rate of 91.7% while maintaining noise rejection capability comparable to existing methods — a capability absent in all prior multi-source RAG frameworks.

Future work will explore three directions: (1) extending the framework to other planetary bodies (Moon, Venus, icy moons) and validating the generalizability of the scale-curvature correspondence and conflict triage principles across different observation ecosystems; (2) incorporating multi-modal retrieval that directly reasons over raw imagery and spectral data rather than metadata-derived knowledge graphs, leveraging vision-language models for planetary scene understanding; and (3) developing an interactive planetary data exploration system that integrates AreoRAG with GIS visualization, enabling scientists to conduct natural language-driven, conflict-aware, multi-scale spatial analysis over the full planetary data archive.

Acknowledgments

This work is supported by the National Key R&D Program of China "Intergovernmental International Science and Technology Innovation Cooperation" (Grant No.2025YFE0107100).

REFERENCES

[1] A. McEwen, S. Byrne, C. Hansen, I. Daubar, S. Sutton, C. Dundas, N. Bardabelias, N. Baugh, J. Bergstrom, R. Beyer, K. Block, V. Bray, J. Bridges, M. Chojnacki, S. Conway, W. Delamere, T. Ebben, A. Espinosa, A. Fennema, J. Grant, V. Gulick, K. Herkenhoff, R. Heyd, R. Leis, L. Ojha, S. Papendick, C. Schaller, N. Thomas, L. Tornabene, C. Weitz, and S. Wilson, "The high-resolution imaging science experiment (hirise) in the mro extended science phases (2009–2023)," *Icarus*, vol. 419, p. 115795, 2024.

[2] M. C. Malin, J. F. Bell III, B. A. Cantor, M. A. Caplinger, W. M. Calvin, R. T. Clancy, K. S. Edgett, L. Edwards, R. M. Haberle, P. B. James, S. W. Lee, M. A. Ravine, P. C. Thomas, and M. J. Wolff, "Context camera investigation on board the mars reconnaissance orbiter," *Journal of Geophysical Research: Planets*, vol. 112, no. E5, 2007.

[3] S. Murchie, R. Arvidson, P. Bedini, K. Beisser, J.-P. Bibring, J. Bishop, J. Boldt, P. Cavender, T. Choo, R. T. Clancy, E. H. Darlington, D. Des Marais, R. Espiritu, D. Fort, R. Green, E. Guinness, J. Hayes, C. Hash, K. Heffernan, J. Hemmler, G. Heyler, D. Humm, J. Hutcheson, N. Izenberg, R. Lee, J. Lees, D. Lohr, E. Malaret, T. Martin, J. A. McGovern, P. McGuire, R. Morris, J. Mustard, S. Pelkey, E. Rhodes, M. Robinson, T. Roush, E. Schaefer, G. Seagrave, F. Seelos, P. Silverglate, S. Slavney, M. Smith, W.-J. Shyong, K. Strohbehn, H. Taylor, P. Thompson, B. Tossman, M. Wirzburger, and M. Wolff, "Compact reconnaissance imaging spectrometer for mars (crism) on mars reconnaissance orbiter (mro)," *Journal of Geophysical Research: Planets*, vol. 112, no. E5, 2007.

[4] D. E. Smith, M. T. Zuber, H. V. Frey, J. B. Garvin, J. W. Head, D. O. Muhleman, G. H. Pettengill, R. J. Phillips, S. C. Solomon, H. J. Zwally, W. B. Banerdt, T. C. Duxbury, M. P. Golombek, F. G. Lemoine, G. A. Neumann, D. D. Rowlands, O. Aharonson, P. G. Ford, A. B. Ivanov, C. L. Johnson, P. J. McGovern, J. B. Abshire, R. S. Afzal, and X. Sun, "Mars orbiter laser altimeter: Experiment summary after the first year of global mapping of mars," *Journal of Geophysical Research: Planets*, vol. 106, no. E10, pp. 23 689–23 722, 2001.

[5] J. P. Grotzinger, J. Crisp, A. R. Vasavada, R. C. Anderson, C. J. Baker, R. Barry, D. F. Blake, P. Conrad, K. S. Edgett, B. Ferdowski, R. Gellert, J. B. Gilbert, M. Golombek, J. Gómez-Elvira, D. M. Hassler, L. Jandura, M. Litvak, P. Mahaffy, J. Maki, M. Meyer, M. C. Malin, I. Mitrofanov, J. J. Simmonds, D. Vaniman, R. V. Welch, and R. C. Wiens, "Mars science laboratory mission and science investigation," *Space Science Reviews*, vol. 170, no. 1, pp. 5–56, 2012.

[6] C. Li, R. Zhang, D. Yu, G. Dong, J. Liu, Y. Geng, Z. Sun, W. Yan, X. Ren, Y. Su, W. Zuo, T. Zhang, J. Cao, G. Fang, J. Yang, R. Shu, Y. Lin, Y. Zou, D. Liu, B. Liu, D. Kong, X. Zhu, and Z. Ouyang, "China's mars exploration mission and science investigation," *Space Science Reviews*, vol. 217, no. 4, p. 57, 2021.

[7] S. Wang, Y. Wang, and H. Wei, "Marsretrieval: Benchmarking vision-language models for planetary-scale geospatial retrieval on mars," *arXiv preprint arXiv:2602.13961*, 2026.

[8] W. Cai, J. Jiang, F. Wang, J. Tang, S. Kim, and J. Huang, "A survey on mixture of experts in large language models," *IEEE Transactions on Knowledge and Data Engineering*, vol. 37, no. 7, pp. 3896–3915, 2025.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive nlp tasks," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS '20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[10] Y. Zhou, Z. Liu, J. Jin, J.-Y. Nie, and Z. Dou, "Metacognitive retrieval-augmented large language models," in *Proceedings of the ACM Web Conference 2024*, ser. WWW '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 1453–1463.

[11] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu, "Unifying large language models and knowledge graphs: A roadmap," *IEEE Trans. on Knowl. and Data Eng.*, vol. 36, no. 7, p. 3580–3599, Jul. 2024.

[12] W. Wu, H. Wang, B. Li, P. Huang, X. Zhao, and L. Liang, "Multirag: A knowledge-guided framework for mitigating hallucination in multi-source retrieval augmented generation," in *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, 2025, pp. 3070–3083.

[13] F. Wang, X. Wan, R. Sun, J. Chen, and S. O. Arik, "Astute RAG: Overcoming imperfect retrieval augmentation and knowledge conflicts for large language models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 30 553–30 571.