TELEHYPERVIS: HYPERGRAPH VISUALIZATION ENHANCED MULTI-MODAL RETRIEVAL-AUGMENTED GENERATION FOR TELECOM STANDARD QA

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) improves the performance of large language models in telecommunications question answering, but existing methods primarily rely on flat vector indices or binary graph structures, limiting their ability to represent higher-order associations among telecom entities. Moreover, retrieved structures are rarely presented in a form that can be jointly interpreted by users and vision-language models. We propose TeleHyperVis-RAG, a multimodal RAG framework that integrates a telecom knowledge corpus, entity–hyperedge retrieval, and multi-view hypergraph visualization. The framework employs dual entity and relation indices, configurable breadth-first neighborhood expansion, and a bipartite-plus-region collage to represent retrieved higher-order relations. We evaluate TeleHyperVis-RAG on five external telecom QA benchmarks. The results show that the proposed configuration achieves the highest observed answer accuracy among the evaluated methods, while latency effects vary across tasks. These findings provide empirical support for combining higher-order retrieval with semantically aligned hypergraph visualization in telecommunications question answering.

1 INTRODUCTION

Large language models (LLMs) have achieved remarkable progress in natural language processing, yet applying them to telecommunications remains challenging due to dense terminology, rapidly evolving standards, and limited parametric knowledge (Maatouk et al. 2024). Retrieval-augmented generation (RAG) mitigates these issues by dynamically incorporating external knowledge, improving accuracy in telecom tasks.

Recent RAG frameworks for telecom have shown promise via hybrid retrieval with 3GPP documents and web search, cross-modal fusion for semantic communication, and multi-modal sensor perception for wireless environments (Bornea et al. 2025; Wang et al. 2025a; Tang et al. 2026; Fan et al. 2025; Mohsin et al. 2025). Despite these advances, existing methods rely on flat vector indices or simple graph structures, failing to capture high-order multi-hop associations and community-level relationships. Moreover, loading and traversing large knowledge structures may introduce additional memory and inference overhead, motivating retrieval architectures that access only query-relevant local subgraphs.

To address multi-hop reasoning bottlenecks, hypergraph-based RAG has been explored. Approaches such as EHRAG, HyperGraphPro, and Cog-RAG leverage hyperedges to model n-ary relations and improve semantic hierarchy understanding (Song et al. 2026; Park et al. 2026; Hu et al. 2025). However, they often lose fine-grained information in large-scale documents, struggle to balance macro-level topics with micro-level details, and incur additional computational costs.

Consequently, two questions arise: (1) *Can a modality-derived telecom knowledge corpus and hypergraph representation support efficient retrieval of higher-order associations?* (2) *Can hypergraph visualization improve downstream QA performance by presenting retrieved structures in a semantically meaningful form, and what accuracy–latency trade-offs does it introduce?*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To answer them, we propose **TeleHyperVis-RAG**, a framework that systematically enhances RAG for telecom multi-modal QA. We build a multi-modal knowledge base covering five major telecom domains via adaptive data sampling, providing cross-modal support for text, time-series signals, and structured metadata. We design an entity-hyperedge binary hypergraph with dual-index storage and a configurable BFS expansion mechanism to capture multi-hop associations. We also introduce a dual-graph collage mechanism combining bipartite incidence and region-selective hypergraph visualizations, achieving multi-scale semantic alignment from macro topics to micro details. Experiments on five telecom QA benchmarks show that TeleHyperVis-RAG achieves the highest observed answer accuracy among the evaluated configurations. The latency results are task-dependent: the framework is competitive on several tasks but incurs additional overhead on some structured reasoning tasks. A paired counterfactual experiment further provides controlled evidence that correct entity-to-structure alignment contributes to the observed visualization benefit.

Our main contributions are summarized as follows.

- We construct an external telecom retrieval corpus from five categories of heterogeneous sources using stratified and topic-diversity sampling. The corpus is independent of the five evaluation benchmarks.
- We propose TeleHyperVis-RAG, combining dual-index hypergraph retrieval, neighborhood expansion, and multi-view visualization to represent n-ary telecom relations at complementary structural scales.
- We evaluate the framework on five external QA benchmarks and multiple generators. Ablation and paired counterfactual experiments provide empirical evidence for higher-order retrieval and semantically aligned visualization.

2 Related Work

2.1 Retrieval-Augmented Generation in Telecommunications

The integration of RAG with telecommunications has progressed along two main directions: hybrid retrieval and multi-modal fusion.

Early hybrid retrieval methods combine 3GPP documents with web search via neural routers (Bornea et al. 2025), but remain text-only and lack adaptability. To address this, multi-modal approaches incorporate RF spectrograms, GPS, LiDAR, camera data, and charts (Khan et al. 2026; Ghosh et al. 2025; Nazar et al., n.d.), enabling wireless environment perception and semantic communication. However, most still rely on flat indices or simple graphs, covering only a few modalities and leaving many telecom subdomains unexplored. We therefore construct a telecom retrieval corpus from five categories of heterogeneous sources, including textual documents, tables and charts, time-series measurements, structured metadata, and other modality-derived records. Non-text sources are converted into textual descriptions for unified indexing while retaining source-specific metadata.

2.2 Hypergraph-Enhanced RAG

Hypergraph-based RAG advances beyond standard entity-relation graphs, which only capture binary associations, by using hyperedges that connect arbitrary nodes to model n-ary relations. The first hypergraph RAG framework adopts a three-stage pipeline: construction, retrieval, and generation (Luo et al. 2025). Later works improve static planning via context-aware decomposition (Zai et al. 2025) or incorporate interaction order as sequence evidence (Wu et al. 2026). These methods enhance multi-hop reasoning, yet they still lose fine-grained details at scale and struggle to balance macro–micro semantics without extra cost.

2.3 Visualization in RAG

Visualization in RAG aids both debugging and model understanding. Early interactive tools help trace retrieval paths (Wang et al. 2025b), while later works use multi-view semantic decoupling (Liu 2025) or multi-granularity multi-modal graphs (Dai et al. 2026) to boost interpretability. However, a single visualization cannot simultaneously reveal macro clusters and micro associations. To address

this, we introduce a dual-graph collage mechanism that combines an incidence bipartite hypergraph and a region-selective hypergraph. This design presents incidence structure and higher-order membership in complementary views. We empirically evaluate whether this multi-view representation improves downstream QA performance and characterize its task-dependent latency overhead.

3 METHODOLOGY

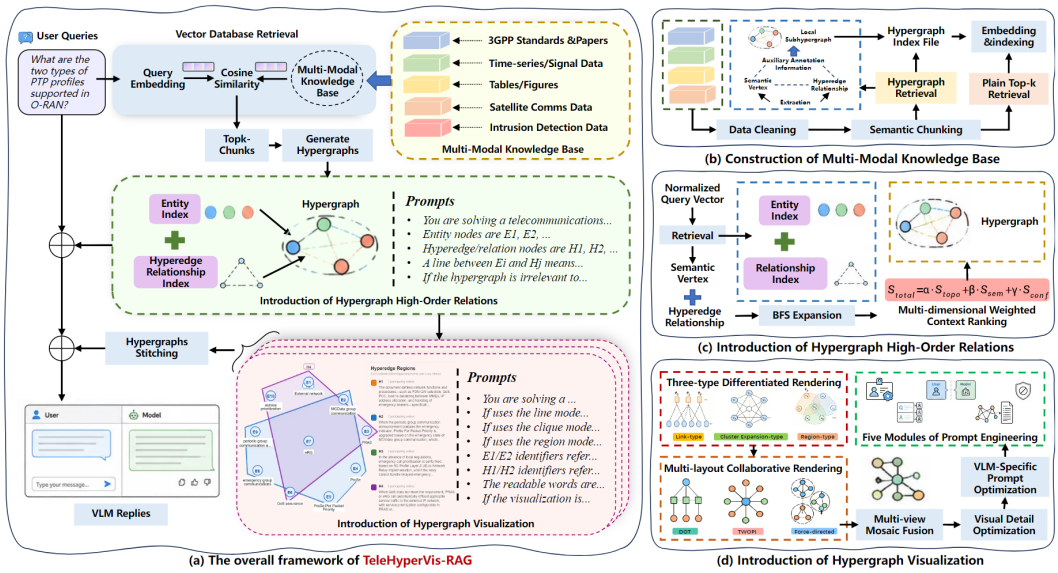


Figure 1: Overall framework of TeleHyperVis-RAG. (a) The overall framework integrating multi-modal knowledge base, hypergraph high-order retrieval, and hypergraph visualization modules. (b) Construction of the multi-modal knowledge base: data cleaning, semantic chunking, embedding vectorization, and hypergraph construction. (c) Hypergraph high-order relations: entity-hyperedge binary modeling, dual-index storage, BFS-based neighborhood expansion, multi-dimensional weighted ranking, and deduplication. (d) Hypergraph visualization: three rendering modes, multi-layout engines, and dual-graph collage mechanism.

3.1 CONSTRUCTION OF MULTI-MODAL KNOWLEDGE BASE

To support complex QA tasks in the telecom domain, we construct a multi-modal knowledge base covering diverse areas including 5G/6G core networks, satellite communications, network security, IoT, and academic standards. Raw multi-modal data sources — ranging from 3GPP specifications and academic papers to RF spectrograms, GPS logs, LiDAR point clouds, and camera images — are first systematically cleaned. Textual documents are directly used without further cleaning, while non-text modalities such as images, charts, and time-series signals are converted into natural language descriptions via a vision-language model, enabling unified semantic processing. The cleaned data are then partitioned into fixed-size semantic chunks, with controlled overlap to preserve boundary continuity. After chunking, two embedding pipelines are adopted: a plain Top-k retrieval path that directly encodes chunks into dense vectors using a multi-modal embedding model, and a hypergraph retrieval path that incorporates a two-stage progressive sampling strategy (stratified random sampling followed by topic diversity sampling) and builds entity-hyperedge hypergraphs from the sampled corpus. Hypergraph elements are extracted using a lightweight LLM, and the resulting hypergraph index is fed into the same embedding model to produce vector representations for high-order relation retrieval. Detailed parameter settings and algorithmic formulations are provided in Appendix.

3.2 Introduction of Hypergraph High-Order Relations

A hypergraph consists of semantic vertices and hyperedges. Each vertex represents a knowledge unit with a unique identifier, type label, description, and source chunk IDs. Each hyperedge connects an arbitrary number of vertices and includes associated entities, relation description, keywords, and source chunk information. To balance expressive power and retrieval efficiency, we adopt a hybrid storage architecture combining a JSON topology file and vector indices. The topology layer stores entity attributes, hyperedge definitions, and bidirectional mappings. The vector layer builds FAISS[1] indices for entities and hyperedge relations using a multi-modal embedding model. This design supports both structural multi-hop reasoning and fast semantic pre-filtering.

Unified Semantic Embedding Vectorization. Given a user query q, we convert it into a dense vector using the embedding model and apply L2 normalization. An exponential backoff retry mechanism ensures robustness; after consecutive failures, a zero vector is returned.

Entity-Relation Dual-Index Independent Retrieval. We search both the entity and hyperedge relation indices to retrieve the top-K_e entities and top-K_r hyperedges (default 5) most similar to the query vector. This dual-index approach captures both entity-level and relation-level semantics, covering the user intent comprehensively. Results include cosine similarity scores and metadata for subsequent expansion.

Configurable-Depth BFS Neighborhood Expansion. Starting from the retrieved core entities, we perform BFS expansion to discover indirectly related knowledge. Let the expansion depth be D. The text chunks obtained at layer d are:

$$\mathcal{C}^{(0)} = \bigcup_{e \in \mathcal{E}_{\text{seed}}} \text{chunk}(e),$$

$$\mathcal{C}^{(d+1)} = \bigcup_{v \in \mathcal{V}^{(d)}} \bigcup_{e \in \Psi(v)} \text{chunk}(e),$$

$$\mathcal{V}^{(d+1)} = \bigcup_{v \in \mathcal{V}^{(d)}} \bigcup_{e \in \Psi(v)} (e \setminus \{v\}),$$

where $\Psi(v)$ is the set of hyperedges associated with vertex v, and $\text{chunk}(e)$ returns the corresponding text chunk. The process iterates layer by layer until depth D. The final candidate set is the union of all layers. Depth D controls the reasoning range: smaller values yield faster responses; larger values uncover deeper associations.

Multi-Signal Candidate Ranking. Candidates are ranked by combining source priority with lexical, relational, and structural relevance: $S(c) = S_{\text{base}}(c) + \lambda_q O_q(c) + \lambda_r O_r(c) + \lambda_s I_s(c)$, where O_q measures query–evidence overlap, O_r captures relation-keyword matching, and I_s aggregates direct matches to retrieved entities, hyperedges, and source chunks. S_{base} prioritizes directly retrieved relation evidence over neighborhood and BFS-expanded evidence. The weights are fixed across compared methods.

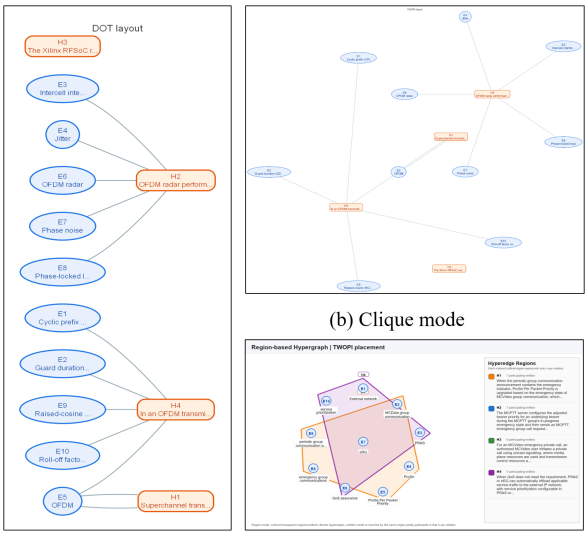
Deduplication and Context Selection. Candidates are sorted by their ranking scores and deduplicated using the first 200 characters of the formatted evidence text as a lightweight content key. The highest-ranked unique candidates are retained until the task-specific context budget K_{final} is reached. If insufficient source-text evidence is available, retrieved entity descriptions are used as a fallback.

3.3 Introduction of Hypergraph Visualization

We design a multi-mode, multi-layout, multi-view hypergraph visualization scheme that integrates rendering algorithms, collage synthesis, and prompt engineering, supporting both human analysis and VLM-based reasoning.

[1] https://github.com/facebookresearch/faiss

Three Differentiated Rendering Modes. We adopt three switchable modes for diverse hyperedge associations. **Incidence mode** (Figure2(a)) constructs a bipartite graph connecting entities to hyperedges via vector lines, with line width proportional to hyperedge weight, suitable for basic association queries. **Clique mode** (Figure2(b)) decomposes hyperedges with ≥ 4 entities into pairwise connections, masking hyperedge nodes to reduce clutter, suitable for lightweight display. **Region mode** (Figure2(c)) uses an improved convex hull algorithm with expansion coefficient (default 1.2) and rounding curvature (default 0.35) to generate semi-transparent colored regions covering associated entities. Unique colors and alpha blending preserve overlapping boundaries, avoiding occlusion and accurately representing high-order relations.



(a) Incidence mode (b) Clique mode (c) Region mode

Figure 2: Three Differentiated Rendering Modes.

Multi-Layout Engine Collaborative Rendering. We integrate three Graphviz layout engines for different topological perspectives: 1) **DOT hierarchical layout** implements the Sugiyama framework, arranging nodes in top-down layers to show vertical relationships such as subordination, suitable for telecom service hierarchies. 2) **TWOPI radial layout** places the highest-degree entity at the center and distributes others on concentric rings according to graph distance (base radius 100px, gradient 50px), highlighting core entities and their influence range. 3) **Force-directed layout** simulates a spring-electrical model (repulsion 0.8, attraction 0.2, 50 iterations), clustering highly related entities and dispersing loosely connected ones, revealing implicit community structures.

Multi-View Collage Fusion Technology. We design an automatic collage mechanism that splices multiple sub-graphs into a single composite image. A unified hash mapping ensures consistent node identification across views. The layout adaptively computes row/column arrangement, trims white borders, and resizes sub-graphs via bilinear interpolation. Each sub-graph is labeled with a descriptive title (e.g., "DOT · incidence"), aiding both human and model interpretation.

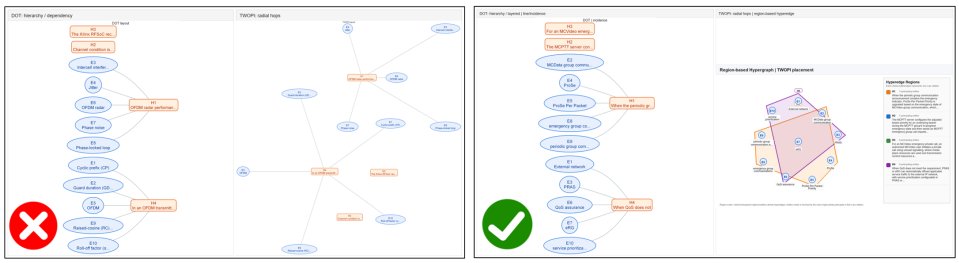


Figure 3: Multi-view collage: region-type plus bipartite graph outperforms bipartite-only collage.

Visualization Parameter Sensitivity. We examine the sensitivity of the visualization to region expansion, transparency, line width, layout engine, and rendering mode. Because no independent development QA set is available, these comparisons are treated as post-hoc sensitivity analyses rather than unbiased hyperparameter optimization. The configuration used in the main experiments employs a region expansion coefficient of 1.2, an alpha value of 0.45, and a line width of 1.5 pixels. These settings produced the highest observed accuracy among the evaluated configurations, but they should not be interpreted as universally optimal.

4 Experimental Results

This section evaluates TeleHyperVis-RAG on five external telecom QA benchmarks: ORAN-Bench-13K text (Gajjar et al. 2025), TeleTable (Zhang et al. 2025), TeleQnA text (Maatouk et al. 2026), TeleMATH (Colle et al., n.d.), and MM-TelcoBench multi-modal (Maatouk et al. 2025). These benchmarks are used as evaluation query sets and are distinct from the external telecom corpus used to construct the retrieval knowledge base. For multiple-choice tasks, we report answer accuracy and average end-to-end inference time. For TeleTable, we additionally report ICC, NAC, and F1 in appendix.

4.1 Experimental Setup and Baseline Definitions

1) Experimental Environment: All experiments are conducted on a unified hardware platform: an Intel® Core™ i9-14900HX processor, 32 GB DDR5 RAM, an NVIDIA GeForce RTX 4060 Laptop GPU (8 GB GDDR6), and Windows 11. The software stack is built on Python 3.10, leveraging libraries such as h5py, json, os, csv, and re for multi-modal data parsing and metric computation. To handle large-scale data, all preprocessing adopts a streaming architecture, with memory footprint per operation controlled within 80 MB.

2) Retrieval Knowledge Corpus and Test Benchmarks: We distinguish the retrieval knowledge corpus from the QA evaluation benchmarks. The retrieval corpus consists of five categories of external telecom data. These data are used exclusively to construct the external RAG knowledge base and are not treated as training, validation, or test samples.The raw knowledge sources include 3GPP specifications, academic papers, open-source datasets, technical metadata, images, charts, and experimental records in formats such as .mat, .raw, .csv, and .db. They are converted into structured JSON retrieval units through the following pipeline: (i) parsing the source files and extracting telecom-specific metadata, such as sampling rate, bandwidth, protocol, and scenario labels; (ii) semantic chunking with a chunk size of 500 tokens and an overlap of 25 tokens; and (iii) streaming the processed chunks to disk for index construction.The QA evaluation data are independent of the retrieval corpus. We evaluate the system on five external telecom QA benchmarks: ORAN-Bench-13K, TeleQnA, MM-TelcoBench, TeleMATH, and TeleTable. The benchmark questions, candidate options, reference answers, and benchmark-specific explanations are not included in the retrieval knowledge base. We evaluate each method on fixed subsets drawn from the official benchmark evaluation splits. All methods compared within the same task use the same evaluation questions.

3) Evaluation Metrics: For multiple-choice tasks, accuracy and average response time are adopted as core metrics. For the TeleTable text-based QA task, we employ key point coverage (ICC), numerical accuracy (NAC), F1 score, accuracy, and average response time for comprehensive assessment. All metrics are computed using a unified evaluation codebase to ensure reproducibility.

4) Baseline Definitions: To systematically validate the effectiveness of each component in our proposed TeleHyperVis-RAG,we define eight comparison systems and the proposed TeleHyperVis-RAG configuration: **a. Base-LLM:** bare large model without retrieval enhancement, relying solely on the model's internal knowledge. **b. Telco-RAG:** plain FAISS vector retrieval on pure 3GPP text, without hypergraph or visualization (Bornea et al. 2025). **c. GraphRAG:** graph-based RAG that constructs knowledge graphs from documents for retrieval (Edge et al. 2024). **d. LightRAG:** dual-level graph retrieval system combining graph structures with vector representations (Guo et al. 2025). **e. HyperGraphRAG:** hypergraph RAG without visualization rendering, only hypergraph-based retrieval (Luo et al. 2025). **f. BiGraph-RAG:** traditional single bipartite hypergraph RAG without region collage visualization optimization. **g. MultiHyper-RAG:** integrates a multi-modal knowledge base, without hypergraph and visualization. **h. TeleHyperVis-RAG:** The final solution

of this work, which integrates a multi-modal knowledge base, hypergraph-based high-order retrieval, and bipartite-and-region graph collage visualization.

Unless otherwise specified, TeleHyperVis-RAG uses the Doubao-embedding-vision-251215 multi-modal embedding model, the task-specific retrieval configurations reported in the experimental tables, and the bipartite+region collage visualization structure, with GPT-5.2 as the generator.

4.2 Performance Gain of Hypergraph Visualization

We first present the core comparison results. Using HyperGraphRAG and BiGraph-RAG as baselines, we compare three configurations: no visualization, single bipartite visualization, and bipartite+region collage visualization, as summarized in Table 1.

The bipartite+region collage visualization achieves the optimal performance, improving accuracy by 2.17% over the no-visualization hypergraph baseline. The single bipartite visualization by contrast degrades performance. On structured table reasoning tasks, visualization helps the model quickly identify entity associations and relational constraints.

Table 1: QA accuracy and end-to-end latency under different retrieval and visualization configurations

Task	Method	Visualization form	Entity K	Relation K	BFS depth	Final K	Accuracy	Avg time
ORAN-Bench-13K	LightRAG	No visualization	-	-	-	10	74%	5.39s
	GraphRAG	No visualization	-	-	-	10	75%	2.94s
	HyperGraphRAG	No visualization	4	3	1	10	78%	13.24s
	BiGraph-RAG	Single bipartite	4	3	1	10	79%	30.93s
	TeleHyperVis-RAG	Bipartite + region collage	4	3	1	10	**81%**	10.99s
TeleQnA	LightRAG	No visualization	-	-	-	15	82%	8.09s
	GraphRAG	No visualization	-	-	-	15	79%	4.27s
	HyperGraphRAG	No visualization	6	6	1	15	83%	8.35s
	BiGraph-RAG	Single bipartite	6	6	1	15	85%	10.81s
	TeleHyperVis-RAG	Bipartite + region collage	6	6	1	15	**87%**	9.22s
MM-TelcoBench	LightRAG	No visualization	-	-	-	8	45%	15.63s
	HyperGraphRAG	No visualization	4	2	1	8	40%	27.95s
	BiGraph-RAG	Single bipartite	4	2	1	8	42%	24.63s
	TeleHyperVis-RAG	Bipartite + region collage	4	2	1	8	**45%**	14.06s
TeleMATH	LightRAG	No visualization	-	-	-	12	25%	5.71s
	GraphRAG	No visualization	-	-	-	12	30%	3.80s
	HyperGraphRAG	No visualization	3	3	1	12	27%	9.84s
	BiGraph-RAG	Single bipartite	3	3	1	12	30%	10.84s
	TeleHyperVis-RAG	Bipartite + region collage	3	3	1	12	**37%**	14.12s
TeleTable	LightRAG	No visualization	-	-	-	6	44.93%	20.27s
	GraphRAG	No visualization	-	-	-	6	44.31%	15.69s
	HyperGraphRAG	No visualization	2	3	1	6	42.51%	19.89s
	BiGraph-RAG	Single bipartite	2	3	1	6	37.68%	32.95s
	TeleHyperVis-RAG	Bipartite + region collage	2	3	1	6	**46.36%**	50.54s

4.3 Post-hoc Parameter Sensitivity Analysis

The five QA benchmarks are used as evaluation sets, whereas the separate telecom corpus is used to construct the RAG knowledge base. Since no independent development QA set is available, the results in this section are treated as post-hoc sensitivity analyses rather than independent hyperparameter optimization.

1) Hypergraph retrieval. We compare Entity K, Relation K, BFS depth, and Final K against StandardRAG, as shown in Table 2. The results are task-dependent: hypergraph retrieval improves the highest observed accuracy on ORAN-Bench-13K, MM-TelcoBench, TeleMATH, and TeleTable, but may introduce additional latency, especially on multimodal and mathematical tasks.

Overall, the results reveal a task-dependent trade-off among retrieval breadth, context size, answer accuracy, and latency. Larger neighborhoods can expose additional evidence but may also introduce redundant or irrelevant context. Since the same QA benchmarks are used for this post-hoc analysis and final reporting, these results should be interpreted as sensitivity observations rather than independent model-selection evidence.

2) Visualization layout. We compare DOT, TWOPI, and FDP layouts with incidence, clique, and region rendering modes. TWOPI and bipartite-plus-region collage achieve the highest observed

Table 2: Hypergraph retrieval parameter comparison across different datasets

Dataset	Entity K	Relation K	BFS depth	Final K	Accuracy	Avg time
ORAN-Bench-13K	–	–	–	5	73%	6.46s
	5	5	2	8	76%	5.25s
	5	5	1	8	79%	5.15s
	5	5	2	12	**80%**	**4.78s**
TeleQnA	–	–	–	5	83%	13.85s
	3	3	1	12	76%	7.63s
	5	5	2	12	78%	**5.27s**
	2	2	1	5	68%	6.72s
	2	2	2	5	83%	7.40s
	2	2	2	10	**84%**	7.23s
MM-TelcoBench	–	–	–	3	33%	**3.87s**
	2	3	1	5	**43%**	26.57s
	2	5	1	8	38%	27.95s
TeleMATH	–	–	–	3	26%	**6.76s**
	2	5	1	12	27%	9.84s
	3	3	2	15	25%	10.04s
	3	3	1	12	**37%**	14.12s
TeleTable	–	–	–	3	40.62%	20.87s
	3	3	1	12	**42.51%**	**19.89s**
	3	3	1	5	40.70%	20.92s
	2	3	1	6	37.68%	32.95s

accuracy in the evaluated configurations, whereas dense clique rendering may reduce performance when many hyperedges are present. These findings indicate task- and graph-density-dependent visualization effects rather than a universally optimal layout.

4.4 Multimodal Corpus Ablation

To examine the contribution of modality-derived knowledge, we compare four corpus configurations: no external knowledge, 3GPP text only, 3GPP plus arXiv text, and the full heterogeneous corpus containing text, tables, modality-derived descriptions, and experimental records. The number of retrieved contexts is fixed at $K = 3$ for all configurations.

The full corpus achieves the highest observed accuracy on the multimodal task and provides smaller gains on several text-oriented tasks. These results suggest that complementary information from tables, images, and measurement records can help answer parameter- and chart-related questions. The latency increase remains moderate in most evaluated settings, although its magnitude varies by task, and the results are presented in Figure 4.

Multi-modal datasets significantly outperform pure text: +5% accuracy on multi-modal tasks, and also provide small gains on text-only tasks. The fusion of text, tables, images, and time-series signals compensates for missing information in single modalities, especially for parameter queries and chart interpretation. The multi-modal configuration increases latency by less than 10%, offering better cost-effectiveness.

4.5 Performance of Hypergraph Structure

The right panel of Figure 4 compares text-based retrieval with HyperGraphRAG on representative telecom tasks. Hypergraph retrieval provides its largest observed gains on tasks requiring multi-entity constraints or structured associations, including ORAN-Bench-13K and selected multimodal and mathematical settings. However, the effect is sensitive to the neighborhood depth and context budget, and several TeleQnA configurations underperform the text-retrieval baseline [2].

In terms of accuracy, pure hypergraph retrieval improves accuracy by 7–7.5 percentage points on text-based tasks, effectively capturing constraint logic among communication parameters. Substantial accuracy gains are also observed on the MM-TelcoBench multi-modal task, while modest improvements are achieved on mathematical and tabular reasoning tasks, confirming the general applicability of hypergraph retrieval to structured logical reasoning tasks.

[2]Effects of ablating visualization from TeleHypervis RAG is validated in Table 1.

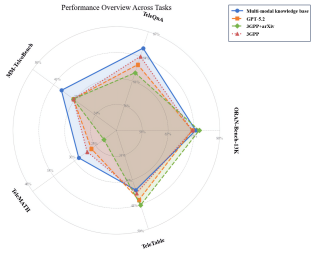


Figure 4: Multi-modal knowledge base ablation study across different tasks (retrieved contexts K = 3).

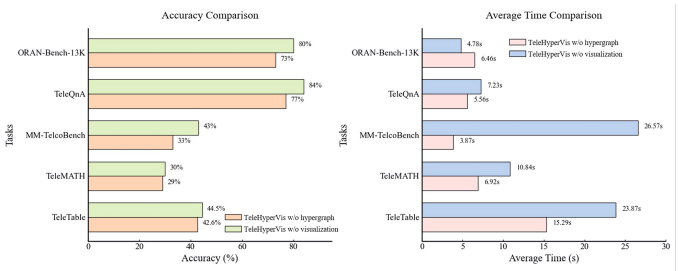


Figure 5: Performance comparison between TeleHyperVis with hypergraph (no visualization) and TeleHyperVis without hypergraph (no visualization) across different tasks.

4.6 Model Agnosticism

We fix the retrieval and visualization configurations and evaluate TeleHyperVis-RAG with GPT-5.2, Gemini-2.5-Flash, Claude-Sonnet-4-6, Claude-Haiku-4.5, and Qwen3.6-27B, as shown in Figure 6.

TeleHyperVis-RAG yields positive average gains for all five generators in the evaluated settings, with a mean improvement of 11.6 percentage points over the corresponding baselines. Gemini-2.5-Flash obtains an average gain of 12.4 percentage points across the five tasks. The magnitude of improvement varies across models, and Qwen3.6-27B benefits less from the more complex visual inputs. These results indicate cross-model applicability while also showing that the effectiveness of hypergraph visualization depends on the generator's visual-structural reasoning capability.

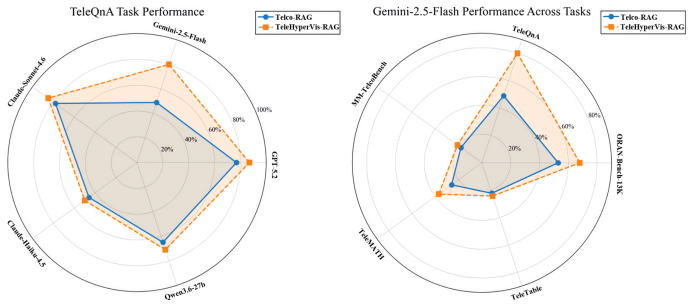


Figure 6: Experimental results demonstrate TeleHyperVis-RAG consistently improves performance across different models.

4.7 Counterfactual Validation

To isolate the semantic contribution of visualization, we conduct a paired counterfactual test on TeleQnA questions. Both conditions share identical retrieved contexts, hypergraph topology, prompts and retrieval parameters; the control only shuffles entity labels while preserving structural links. The correct collage outperforms the shuffled control by 4.0 percentage points (87% vs. 83%), confirming that proper entity-to-structure alignment drives the visualization gain.

5 Conclusion

We present TeleHyperVis-RAG, a multimodal telecom QA framework integrating heterogeneous knowledge, dual-index hypergraph retrieval, neighborhood expansion and multi-view collage visualization. It achieves the highest accuracy across five benchmarks with task-dependent latency, and counterfactual tests verify the benefit of semantically aligned visualization. Future work will focus on native multimodal indexing, efficient rendering and larger-scale validation.

REFERENCES

Alibaba Cloud Tianchi. n.d. low_orbit_satellite_jam_data_100. *https://tianchi.aliyun.com/dataset/198601*.

Bhat, S. R.; et al. 2025. Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis. *arXiv preprint arXiv:2505.21700*.

Boakye-Boateng, K.; Ghorbani, A. A.; Lashkari, A. H. 2023. Securing Substations with Trust, Risk Posture and Multi-Agent Systems: A Comprehensive Approach. In *20th International Conference on Privacy, Security and Trust (PST)*.

Bornea, A.; Ayed, F.; Domenico, A. D.; Piovesan, N.; Salem, T. Si; Maatouk, A. 2025. Telco-oRAG: Optimizing Retrieval-Augmented Generation for Telecom Queries via Hybrid Retrieval and Neural Routing. *IEEE Journal on Selected Areas in Communications*, vol. 44, pp. 2334–2347.

Botez, R.; Zinca, D.; Dobrota, V. 2025. Redefining 6G Network Slicing: AI-Driven Solutions for Future Use Cases. *Electronics*, vol. 14, no. 2, p. 368.

Chouman, A.; Manias, D. M.; Shami, A. 2022. Towards Supporting Intelligence in 5G/6G Core Networks: NWDAF Implementation and Initial Analysis. *arXiv preprint arXiv:2205.15121*.

Coldwell, C.; et al. 2022. Machine Learning 5G Attack Detection in Programmable Logic. In *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 1365–1370.

Colle, V.; Sana, M.; Piovesan, N.; Domenico, A. D.; Ayed, F.; Debbah, M. n.d. TeleMath: A Benchmark for Large Language Models in Telecom Mathematical Problem Solving. *IEEE Network*, to be published.

Dai, S.; Huang, Q.; You, X.; Yu, J. 2026. MG²-RAG: Multi-Granularity Graph for Multimodal Retrieval-Augmented Generation.

Draper-Gil, G.; Lashkari, A. H.; Mamun, M. S. I.; Ghorbani, A. A. 2016. Characterization of Encrypted and VPN Traffic using Time-related Features. In *Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP)*.

Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., & Larson, J. (2024). From Local to Global: A Graph RAG Approach to Query-Focused Summarization. *arXiv preprint arXiv:2404.16130. https://arxiv.org/abs/2404.16130*.

Fan, D.; Meng, R.; Gao, S.; Xu, X. 2025. KGRAG-SC: Knowledge Graph RAG-Assisted Semantic Communication. In *2025 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 1–7.

Feng, Y.; et al. 2022. DeepSense 6G: A Large-Scale Multi-Modal Dataset for Next-Generation Wireless Systems. *arXiv preprint arXiv:2206.09174*.

Feng, A.; et al. 2025. TelecomTS: A Multi-Modal Observability Dataset for Time Series and Language Analysis. *arXiv preprint arXiv:2510.06063*.

Gajjar, P.; Shah, V. K. 2025. ORAN-Bench-13K: An Open Source Benchmark for Assessing LLMs in Open Radio Access Networks. In *2025 IEEE 22nd Consumer Communications & Networking Conference (CCNC)*, pp. 1–4.

Ghosh, R.; Liu, C.; Rele, G.; Ravipati, V. S.; Aouad, H. 2025. TelcoAI: Advancing 3GPP Technical Specification Search through Agentic Multi-Modal Retrieval-Augmented Generation. *arXiv preprint arXiv:2601.16984*.

Guo, Z., Xia, L., Yu, Y., Ao, T., & Huang, C. (2025). LightRAG: Simple and Fast Retrieval-Augmented Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025* (pp. 10746–10761), Suzhou, China.

Hamza, A.; et al. 2023. CIC-IoT-2023: A Comprehensive IoT Intrusion Detection Dataset. *IEEE Internet of Things Journal*.

Hu, H.; et al. 2025. Cog-RAG: Cognitive-Inspired Dual-Hypergraph with Theme Alignment Retrieval-Augmented Generation. *AAAI Conference on Artificial Intelligence*.

Khan, M. Zakir; Ge, Y.; Mollel, M.; Mccann, J.; Abbasi, Q. H.; Imran, M. 2026. RFSensingGPT: A Multi-Modal RAG-Enhanced Framework for Integrated Sensing and Communications Intelligence in 6G Networks. *IEEE Transactions on Cognitive Communications and Networking*, vol. 12, pp. 298–311.

Kim, B.; Kim, J.; Kim, H. 2024. WASD: Wireless Anomaly Signal Dataset for Spectrum Anomaly Detection. *IEEE Access*, vol. 12, pp. 168721–168732.

Kulmar, M. 2024. Radio Access Network Subslicing in Closed Control Loop for Improved Slice Performance Management. TalTech Press.

Lensu, U. 2025. Impact of chunking granularity on accuracy and token consumption in retrieval-augmented generation for question-answering. Master's thesis, Aalto University.

Li, Y.; et al. 2022. V2X-Sim: Multi-Agent Collaborative Perception Dataset and Benchmark for Autonomous Driving. *IEEE Robotics and Automation Letters*.

Liu, J. 2025. Think Parallax: Solving Multi-Hop Problems via Multi-View Knowledge-Graph-Based Retrieval-Augmented Generation. *arXiv preprint arXiv:2510.15552*.

Luo, H.; et al. 2025. HyperGraphRAG: Retrieval-Augmented Generation via Hypergraph-Structured Knowledge Representation.

Maatouk, A.; Ampudia, K. C.; Ying, R.; Tassiulas, L. 2024. Tele-LLMs: A Series of Specialized Large Language Models for Telecommunications. *arXiv preprint arXiv:2409.05314*.

Maatouk, A.; Piovesan, N.; Ayed, F.; De Domenico, A.; Debbah, M. 2025. Large Language Models for Telecom: Forthcoming Impact on the Industry. *IEEE Communications Magazine*, vol. 63, no. 1, pp. 62–68.

Maatouk, A.; Ayed, F.; Piovesan, N.; Domenico, A. D.; Debbah, M.; Luo, Z.-Q. 2026. TeleQnA: A Benchmark Dataset to Assess Large Language Models Telecommunications Knowledge. *IEEE Network*, vol. 40, no. 2, pp. 253–260.

Mohsin, M. A.; Bilal, A.; Bhattacharya, S.; Cioffi, J. M. 2025. Retrieval Augmented Generation with Multi-Modal LLM Framework for Wireless Environments. In *2025 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 184–189.

Nazar, A. M.; Celik, A.; Selim, M. Y.; Abdallah, A.; Qiao, D.; Eltawil, A. M. n.d. ENWAR: A RAG-Empowered Multi-Modal LLM Framework for Wireless Environment Perception. *IEEE Communications Magazine*, to be published.

Nugraha, B.; et al. 2025. A Comprehensive 5G Network Dataset for Control and Data Plane Security and Resource Allocation Analysis. In *2025 IEEE International Conference on Cyber Security and Resilience (CSR)*.

O'Shea, T. J.; West, N. 2016. Radio Machine Learning Dataset Generation with GNU Radio. In *Proceedings of the GNU Radio Conference*.

Park, J.; Lee, S.; Khan, O. Z.; Kim, H. J.; Kim, J. 2026. HyperGraphPro: Progress-Aware Reinforcement Learning for Structure-Guided Hypergraph RAG.

Quetschlich, N.; Burgholzer, L.; Wille, R. 2023. MQT Bench: Benchmarking Software and Design Automation Tools for Quantum Computing. *Quantum*.

Saz, A. F.; Xiong, S.; Saidutta, Y. M.; Fekri, F. 2024. Model-Theoretic Logic for Mathematical Theory of Semantic Information and Communication. *arXiv preprint arXiv:2401.17556*.

Sharafaldin, I.; Lashkari, A. H.; Ghorbani, A. A. 2018a. A Detailed Analysis of the CSE-CIC-IDS2018 Dataset. In *2018 IEEE International Conference on Big Data*.

Sharafaldin, I.; Lashkari, A. H.; Ghorbani, A. A. 2018b. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*.

Song, Y.; Tao, X.; Yang, Z.; Luo, Y.; Tang, J. 2026. EHRAG: Bridging Semantic Gaps in Lightweight GraphRAG via Hybrid Hypergraph Construction and Retrieval.

Tang, S.; et al. 2026. Retrieval-Augmented Generation for GenAI-Enabled Semantic Communications. *IEEE Wireless Communications*, vol. 33, no. 1, pp. 259–268.

Uhlig, S.; Quoitin, B.; Lepropre, J.; Balon, S. 2006. Providing public intradomain traffic matrices to the research community. *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, pp. 83–86.

Wang, Y.; Zhang, Y.; Xu, K.; Xu, Y.; Wang, Y.; Feng, Q. 2025a. CORB-RAG: A Comprehensive Evaluation Benchmark for Retrieval-Augmented Generation Systems in the Chinese Telecommunications Operator Domain. In *2025 IEEE 5th International Conference on Computer Communication and Artificial Intelligence (CCAI)*, pp. 503–508.

Wang, K.; et al. 2025b. XGraphRAG: Interactive Visual Analysis for Graph-based Retrieval-Augmented Generation. In *2025 IEEE 18th Pacific Visualization Conference (PacificVis)*, pp. 1–11.

Wu, K.; et al. 2026. Knowledge Is Not Static: Order-Aware Hypergraph RAG for Language Models.

Zai, X.; Tan, X.; Wang, X.; Liu, Q.; Xu, X.; Zhang, W. 2025. PRoH: Dynamic Planning and Reasoning over Knowledge Hypergraphs for Retrieval-Augmented Generation. In *Proceedings of the ACM Web Conference 2026*.

Zegura, E. W.; Calvert, K. L.; Bhattacharjee, S. 1996. How to model an internetwork. In *Proceedings of IEEE INFOCOM '96*.

Zhang, J.; et al. 2025. T2R-bench: A benchmark for generating article-level reports from real world industrial tables. *arXiv preprint arXiv:2508.19813*.

Zhao, J.; Pan, J. 2024. LENS: A LEO Satellite Network Measurement Dataset. In *Proceedings of the 15th ACM Multimedia Systems Conference*, pp. 278–284.

A Appendix

A.1 Construction of the External Retrieval Knowledge Base

Retrieval Corpus Preparation. For the communication field, the academic and open-source communities have released authoritative datasets covering multiple dimensions, including 5G/6G core networks and wireless communications, satellite communications and anti-jamming, network security and intrusion detection, IoT and industrial/vehicular communications, academic standards and open-source tools.

Specifically, they include: 1) **5G/6G core network and wireless communication scenario datasets**: 5GDatasets-main (Nugraha et al. 2025), 5G-Core-Networks-Datasets-main (Chouman et al. 2022), 5GAD-main (Coldwell et al. 2022), deepsense6G (Feng et al. 2022), 6g-network-slicing-dataset-main (Botez et al. 2025), nsfnet (Zegura et al. 1996), geant2 (Uhlig et al. 2006), WASD (Kim et al. 2024), Inductive-Semantic-Communication-Dataset-main (Saz et al. 2024). 2) **Satellite communication and anti-jamming / quantum communication datasets**: low_orbit_satellite_jam_data_100 (Alibaba Cloud Tianchi, n.d.), LENS-master (Zhao et al. 2024). 3) **Network security and intrusion detection datasets**: CIC-IDS2017 (Sharafaldin et al. 2018b), CIC-IOT2023 (Hamza et al. 2023), CSE-CIC-IDS2018 (Sharafaldin et al. 2018a), ISCXVPN2016 (Draper-Gil et al. 2016), Modbus-2023 (Boakye-Boateng et al. 2023). 4) **IoT and industrial/vehicular communication datasets**: MCCL_public_dataset (Kulmar 2024), RML2016.10a+04c (O'Shea et al. 2016), TelecomTS (Feng et al. 2025), V2X-Sim-2.0-mini (Li et al. 2022). 5) **Academic standards / open-source tool dataset**: MQTBench (Quetschlich et al. 2023).

These five categories constitute the source corpus used to build the RAG knowledge base; they are not QA training or test datasets. ORAN-Bench-13K, TeleQnA, MM-TelcoBench, TeleMATH, and TeleTable are maintained separately and are used only as external QA evaluation benchmarks.

Data Cleaning. Raw multi-modal data sources are highly heterogeneous and must be systematically cleaned before building a high-quality RAG index. The text data are from 3GPP specification documents and papers retrieved by keywords, with standardized formats that do not require cleaning. For non-text modalities such as images, charts, time-series signal plots, and structured metadata, we use a vision-language model to automatically generate corresponding natural language descriptive annotations, converting them into knowledge units that can be embedded into the vector space. Let the original representation of multi-modal data M be $M \in \{\mathcal{G}, \mathcal{T}, \mathcal{R}\}$, where $\mathcal{G}$ denotes scene images and flowcharts, $\mathcal{T}$ tables, and $\mathcal{R}$ real test data. The description text is generated by the vision-language model f_{VLM} as:

$$\text{Caption}(M) = f_{\text{VLM}}(M; \theta_{\text{VLM}}),$$

where θ_{VLM} are the parameters of the Doubao vision-language model (Doubao-Seed-2.0-lite-260215)[3].

Benchmark Independence and Leakage Control. The retrieval corpus does not contain benchmark questions, candidate options, reference answers, benchmark annotations, or question-specific explanations. It may contain independently authored standards or technical documents that provide the factual evidence needed to answer benchmark questions, which is expected in an open-book RAG setting. We distinguish legitimate evidence overlap from benchmark leakage: independently authored source documents are retained, whereas benchmark-derived question and answer content is excluded.

Text Semantic Chunking. To support efficient indexing and retrieval, long documents are split into appropriately sized semantic units. We adopt a fixed-size chunking strategy, setting chunk size to 500 tokens and overlap to 25 tokens. For a document with total length L tokens, the number of chunks N is calculated as:

$$N = \left\lceil \frac{L - c}{c - o} \right\rceil + 1,$$

[3] https://research.doubao.com/zh/seed2

where $c = 500$, $o = 25$. The overlapping design ensures semantic continuity between adjacent chunks, avoiding the breakage of key information at boundaries. Smaller chunks (64–128 tokens) are suitable for fact-based short answer queries, while larger chunks (512–1024 tokens) are more appropriate for complex tasks requiring extensive context (Bhat et al. 2025; Lensu 2025). Our chosen 500-token chunk size achieves a balance between fine-grained semantic capture and contextual integrity.

Embedding Vectorization. After chunking, each text chunk needs to be converted into a vector embedding for similarity retrieval. We design two different processing pipelines for two types of retrieval tasks.

1) **Plain Top-k Retrieval Embedding:** For ordinary QA queries that do not require complex multi-hop reasoning, we directly encode the text chunks into fixed-dimensional dense vectors using the embedding model(Doubao-embedding-vision-251215)[4]:

$$\mathbf{v}_i = f_{\text{embed}}(\text{chunk}_i) \in \mathbb{R}^d,$$

where $d = 1024$. Given a user query q, its embedding $\mathbf{v}_q$ and the chunk embeddings are compared by cosine similarity:

$$\text{sim}(\mathbf{v}_q, \mathbf{v}_i) = \frac{\mathbf{v}_q \cdot \mathbf{v}_i}{\|\mathbf{v}_q\|\|\mathbf{v}_i\|}.$$

2) **Hypergraph Retrieval Embedding:** For complex telecom standard queries requiring multi-hop associations and community-level relationships, we design a dedicated pipeline with two-level progressive sampling and hypergraph construction.

a. Stratified Random Sampling : For 3GPP specification documents, each specification file must contribute a certain number of samples. Let the set of 3GPP files be $\mathcal{F}_k$, the set of chunks from file f be $\mathcal{C}_f$, and the number of samples extracted from that file be n_f. Under the constraint of a minimum sampling rate α per file, we make the total number of samples as close as possible to the preset quota Q_k:

$$\min \sum_k \left| \sum_{f \in \mathcal{F}_k} n_f - Q_k \right| \quad \text{s.t.} \quad n_f \geq \alpha |\mathcal{C}_f|, \forall f.$$

where Q_k denotes the total number of samples expected to be drawn from the k-th file set. If the quota is not met, additional samples are drawn from large-capacity files. For arXiv academic papers and general knowledge data, since there is no strict file hierarchy, we directly adopt pure random sampling without imposing per-file lower bounds.

b. Topic Diversity Sampling: General knowledge datasets contain many tables, special symbols, and dense numbers, making text features easily distorted and unsuitable for cluster analysis. Therefore, we apply a statistical indicator-based filtering strategy: filter out texts longer than 1024 tokens or shorter than 50 tokens, then compute special character ratio ρ_{spec}, digit density ρ_{digit}, blank line ratio ρ_{blank} for each sample. If $\rho_{\text{spec}} > \theta_{\text{spec}}$ or $\rho_{\text{digit}} > \theta_{\text{digit}}$, the sample is judged as low-quality noise and removed. After filtering, a fixed number Q_{gen} of samples are randomly drawn from the clean texts. The decision function can be expressed as:

$$\text{Keep}(s) = \begin{cases} 1, & \begin{cases} L_{\min} \leq |s| \leq L_{\max} \\ \rho_{\text{spec}}(s) \leq \theta_{\text{spec}} \\ \rho_{\text{digit}}(s) \leq \theta_{\text{digit}} \end{cases} \\ 0, & \text{otherwise.} \end{cases},$$

For well-structured and strongly professional 3GPP and arXiv texts, we proceed to a clustering pipeline. First, TF-IDF is used for text vectorization. Suppose the document set $\mathcal{D}$ contains N text chunks, vocabulary size V. The feature vector $\mathbf{x}_j \in \mathbb{R}^V$ for chunk j has its t-th dimension given by:

$$\text{TF-IDF}(t, j) = \frac{n_{t,j}}{\sum_k n_{k,j}} \cdot \log \frac{N}{|\{j' : t \in j'\}|}.$$

[4]https://seed1-6-embedding.github.io

Then MiniBatch KMeans clustering is applied to partition the texts into semantic topic clusters. The objective is to minimize the sum of squared distances from each sample to its cluster center:

$$\min_{\{\boldsymbol{\mu}_t\}_{t=1}^{K}} \sum_{j=1}^{N} \min_{t} \|\mathbf{x}_j - \boldsymbol{\mu}_t\|^2,$$

where $\boldsymbol{\mu}_t$ is the center of the t-th cluster. To protect niche topics and suppress over-sampling of popular ones, we fully preserve small clusters and impose a sampling cap on large clusters. Let the size of cluster t be $|\mathcal{C}_t|$ and the total sampling quota be Q_{academic}. If $|\mathcal{C}_t| < \tau$, the whole cluster is kept; otherwise, the number of samples drawn from it is:

$$n_t = \min\left(|\mathcal{C}_t|,\ \gamma \cdot \frac{Q_{\text{academic}}}{K}\right),$$

where $\gamma \geq 1$ is a capacity adjustment factor.

After all sampling steps, the two sets are merged and shuffled to produce a high-quality, topic-balanced corpus.

c. Hypergraph Construction and Embedding: After sampling, we use GPT-5.4-Nano [5] to perform semantic decomposition on each text chunk. Standardized prompts are used to give structured parsing instructions to GPT-5.4-Nano, extracting semantic vertices (including protocol terms, technical entities, theoretical concepts, functional modules, performance metrics, standard entries, experimental conclusions, etc.) and multi-level hyperedge relations (subordination, logical deduction, technical constraints, scenario applications, version evolution, principle support, parameter correspondence, etc.), along with auxiliary annotations such as entity domain attributes, hierarchy categories, and data source. After extracting local vertices and relations within a single chunk, a local sub-hypergraph is formed for each chunk. Finally, the vertices and hyperedge attributes are organized according to the storage format and indexing rules of the hypergraph RAG system to generate a hypergraph index file, which is then fed into the Doubao vision-language model for vector embedding to obtain vector representations usable for hypergraph retrieval.

A.2 Modular Prompt Engineering and Prompt Templates

VLM-Specific Prompt Optimization Strategy. We enhance the model's graph reading and reasoning ability from four dimensions: template structure, domain instructions, text-image alignment, and few-shot examples.

1) **Structured hierarchical prompt template:** The prompt is decomposed into four fixed layers: general instructions, graph interpretation, task requirements, and output format. General instructions inform the model to recognize hypergraph characteristics; graph interpretation clearly defines the semantic meanings of nodes, hyperedges, regions, and colors for the three visualization modes; task requirements specify the QA task; output format constrains the final answer presentation.

2) **Hypergraph-specific instruction optimization:** Domain-specific terms such as "high-order hyperedge", "entity co-occurrence", "overlapping association", "topological clustering" are injected, together with telecom domain prior knowledge, to enhance the model's perception of professional graphs. Constraint instructions are also added, requiring the model to ignore invalid pixels and focus on core topological structures.

3) **Text-image alignment prompt concatenation:** The automatically generated visualization graph description text, Base64-encoded image, and standardized interpretation rules are concatenated into a unified multi-modal input. The order of text and image is controlled: visual rules are presented first, then the graph material.

4) **Few-shot examples:** Five standard hypergraph interpretation examples are embedded at the beginning of the prompt, each containing a typical hypergraph visualization image and its corresponding correct answer and brief explanation. This allows the model to quickly adapt to the rendering rules and answering paradigm of our visualization module, significantly reducing cold-start recognition errors.

[5] https://nano-gpt.com/models/text/openai/gpt-5.4-nano

Prompt for the hypergraph-enhanced RAG framework

You are solving a telecommunications multiple-choice question.

You are given:
1. A question with candidate options.
2. Retrieved textual evidence.
3. An image visualizing the relevant local hypergraph.

Instructions:
- Use the question and options as the primary basis for answering.
- Treat the retrieved evidence and hypergraph image as supplementary information; ignore them if irrelevant.
- The image may contain multiple panels. All panels represent the same local hypergraph using different layouts or visual encodings.
- In an incidence view, a connection between an entity node (Ei) and a hyperedge node (Hi) indicates that Ei participates in Hi.
- In a region view, entity nodes enclosed by or touching the same region (Hi) jointly participate in that hyperedge.
- Overlapping regions indicate entities shared by multiple hyperedges.
- Do not interpret different panels as separate hypergraphs.
- Return only the number of the correct option, without any explanation or additional text.

[Question and Options]
{question}

[Retrieved Evidence]
{retrieved_context}

[Hypergraph Visualization]
<image>

Figure 7: Prompt template for the proposed hypergraph visualization-enhanced RAG framework.

A.3 Additional TeleTable Results

TeleTable Parameter and Visualization Sensitivity. We report additional TeleTable results under different hypergraph retrieval parameters and visualization modes. Node K, Relation K, BFS depth, and Final K denote the number of initially retrieved entities, retrieved relations, BFS expansion layers, and final hypergraph contexts, respectively. All configurations additionally use five original table contexts. Since no independent development QA set is available, the results are presented as post-hoc sensitivity analyses.

Table 3: TeleTable results under different retrieval parameters and visualization modes.

Node K	Relation K	BFS depth	Final K	Visualization mode	ICC	F1	Accuracy	Avg. time
2	3	1	6	Bipartite + region collage	**46.49%**	**38.57%**	**46.36%**	50.54s
–	–	–	5	None	43.80%	33.25%	40.70%	20.92s
–	–	1	12	None	42.38%	33.30%	40.54%	21.43s
3	3	1	12	None	46.19%	33.22%	42.51%	**19.89s**
5	5	1	12	None	15.90%	9.65%	17.27%	25.63s
2	3	1	6	Single bipartite	40.63%	30.40%	37.68%	32.95s

Note: All configurations additionally include five original table contexts. "None" indicates that only textual retrieval contexts are provided without graph visualization. A dash indicates that the corresponding parameter was not explicitly recorded in the experiment log. Bold values denote the highest observed score or the lowest average inference time.

Under the same retrieval setting $(2, 3, 1, 6)$, the bipartite-plus-region collage achieves 46.36% accuracy, compared with 37.68% for the single bipartite visualization. It also improves ICC and F1 by 5.86 and 8.17 percentage points, respectively, although its average inference time increases from 32.95 to 50.54 seconds. Among the non-visual configurations, the setting with Node $K = 3$, Re-

lation $K = 3$, and Final $K = 12$ obtains the highest observed accuracy of 42.51% and the lowest average latency of 19.89 seconds. In contrast, retrieving five nodes, five relations, and 12 final contexts substantially reduces performance, suggesting that larger retrieval neighborhoods may introduce irrelevant or redundant evidence. Overall, the results show a trade-off among retrieval scope, visualization design, accuracy, and latency.

A.4 Hypergraph Visualization Cases

Hierarchical Modular Prompt Architecture. Based on the VLM-specific optimization above, we build a hierarchical modular prompt engineering for the whole hypergraph-enhanced retrieval generation process. The complete prompt is divided into five independent modules arranged in logical progression: 1) **Task identity definition module:** informs the model to act as a communications domain expert, clarifying its knowledge background and reasoning responsibility. 2) **Unified hypergraph interpretation guidance module:** provides standardized explanations for the three visualization modes (incidence, clique, region), and clarifies that multiple sub-graphs in the collage belong to the same set of local hypergraphs, eliminating cognitive bias caused by different visualization forms. 3) **QA core content module:** carries the user's original question and candidate options. 4) **Auxiliary knowledge evidence module:** embeds retrieved and filtered textual evidence (e.g., standard clauses, technical specification fragments, experimental conclusions). 5) **Answer behavior constraint module:** strictly restricts the output format, allowing only the final answer index (no additional explanation or reasoning process), and guides the model to ignore graph content irrelevant to the current question.

The information arrangement strictly follows the order "rules first, then text, then image": first, the general hypergraph reading rules and domain answering requirements are passed; then the question, options, and textual evidence are input; finally, the hypergraph visualization image is embedded. In addition, domain reasoning bias guidance is incorporated into the prompt, emphasizing that the model should rely primarily on high-order knowledge represented by the hypergraph (multi-entity collaborative constraints, technical parameter association logic, experimental laws) rather than shallow textual semantic matching. This entire prompt strategy does not rely on temporary wording adjustments but uses a fixed modular structure and unified guidance phrases to achieve standardized reuse, significantly improving the stability and accuracy of model reasoning in complex telecom professional scenarios.

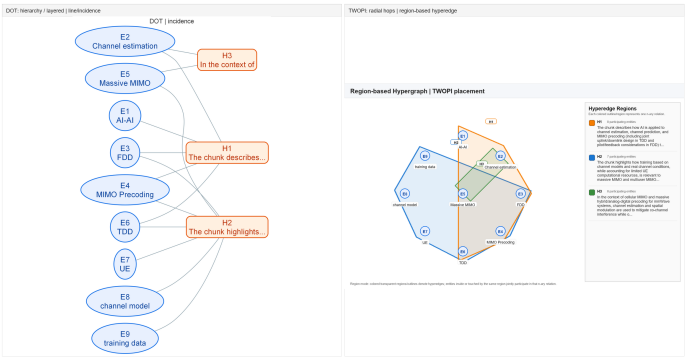
Prompt Templates for RAG Generation. We design a dedicated generation prompt template for the proposed hypergraph visualization-enhanced RAG framework, which is tailored for telecommunications multiple-choice question answering. The template integrates textual evidence and hypergraph visual information, and explicitly specifies the interpretation rules for different visualization views (incidence view and region view) to guide the model to accurately utilize multi-modal retrieval results for reasoning.

We present six groups of general hypergraph visualization cases with different topological scales and structural characteristics. Each case adopts a collage layout of incidence mode and region mode: the incidence mode intuitively presents the bipartite incidence structure between entities and hyperedges, while the region mode displays high-order clustering and overlapping association relationships among entities through region filling and color coding.

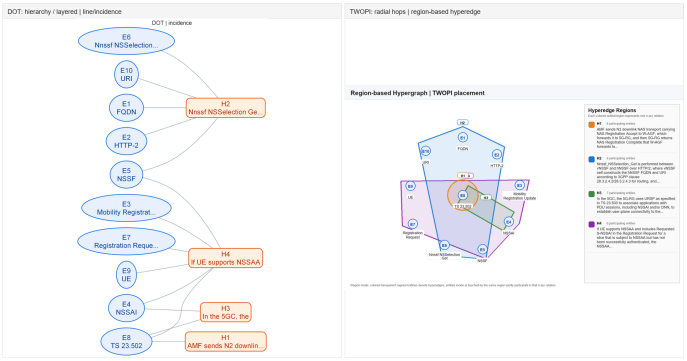
A.5 Failure Cases and Experimental Reflection

We summarize all negative-gain, failure, and anomalous cases: 1) **Knowledge base redundancy negative gain:** On ORAN-Bench-13K and TeleTable, full generic knowledge or excessive tables hurt performance (ORAN-Bench-13K -1%, TeleTable -3.8%), showing that there exists an optimal capacity threshold; indiscriminate data stacking dilutes professional knowledge density and introduces noise interference.

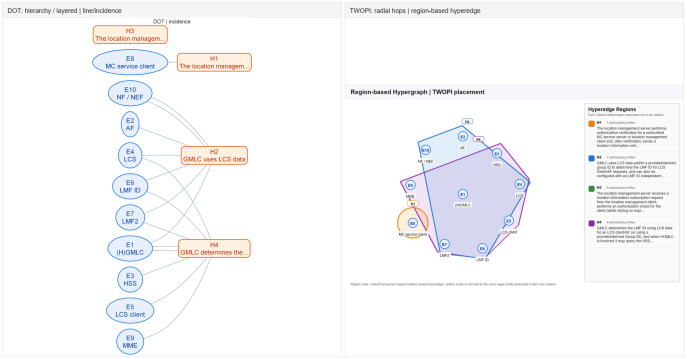
2. **Visualization structure mismatch:** Single bipartite graph reduces accuracy by 4–8% compared to no-visualization hypergraph; clique mode with more than 5 hyperedges causes severe node entanglement (accuracy -6%). Medium reasoning intensity causes overfitting on multi-modal tasks (-3%). 3. **Hyperparameter threshold:** BFS depth=2 is worse than depth=1; large K values also



(a) Visualization case 1

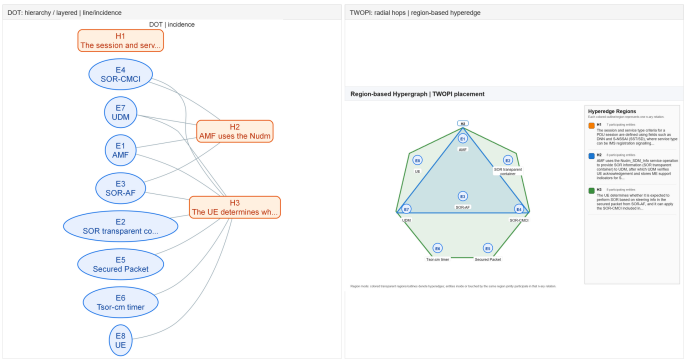


(b) Visualization case 2

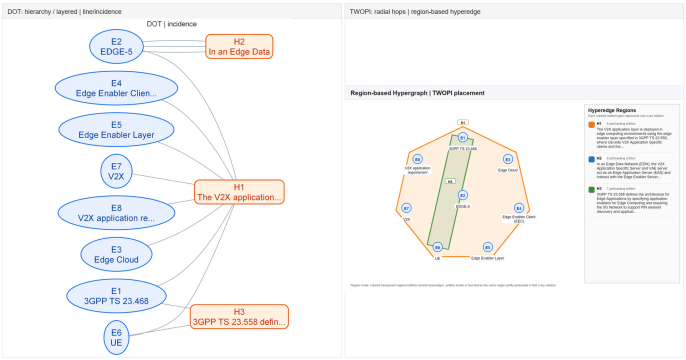


(c) Visualization case 3

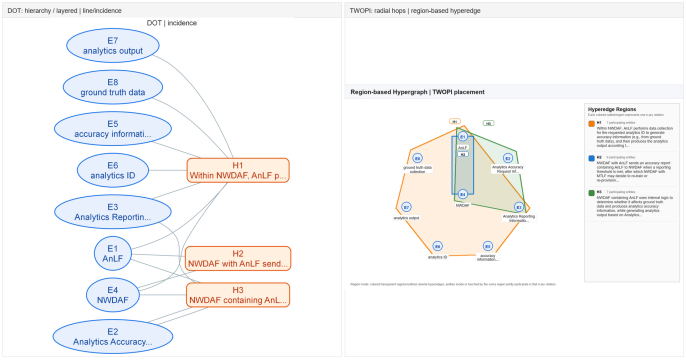
Figure 8: Hypergraph visualization cases (Part 1/2).



(a) Visualization case 4



(b) Visualization case 5



(c) Visualization case 6

Figure 9: Hypergraph visualization cases (Part 2/2, incidence + region collage mode) with different topological structures.

hurt. The combination (8/8/1/10) gives only 69% accuracy (-11% from optimum). Excessive neighborhood expansion introduces irrelevant entities. 4. **Model compatibility boundary:** Qwen3.6-27b achieves only 71% accuracy with high latency, showing that lightweight open-source models have limited capacity for complex hypergraph visualizations.

A.6 LIMITATIONS

The current framework suffers from three main limitations. First, when the number of hyperedges is large, the constructed graph becomes overly dense, which increases the difficulty of model interpretation and may introduce redundant topological noise. Second, the full multi-modal knowledge base introduces redundant noise in some pure text tasks, resulting in unnecessary computational overhead. Third, lightweight open-source models have limited ability to interpret collage-style multi-view visualizations, and the performance gain from visual enhancement is not significant on small models.

Future work will explore adaptive hyperedge pruning and dynamic aggregation mechanisms to alleviate the graph density problem; introduce knowledge base quality assessment and selective sampling strategies to remove redundancy; develop simplified visualization variants and lightweight prompt templates tailored for small models; and study end-to-end latency-accuracy joint optimization methods to promote deployment on resource-constrained edge devices.

A.7 AI USE STATEMENT

We confirm that large language models (LLMs) and generative AI tools were used in this work:

1. To aid and polish the writing of the manuscript;
2. For literature retrieval and discovery of related work.

All AI-assisted content has been manually reviewed and validated by all authors. The authors take full responsibility for the final content of this paper.