SynapseHSG: One-Shot Hyper-Semantic Grounding for Long-Context Multimodal Question Answering

Anonymous ACL submission

Abstract

Answering complex queries over long-context multimodal streams requires dynamically synthesizing asynchronous and heterogeneous evidence across extended temporal and logical spans. Existing approaches often rely on iterative Multimodal Retrieval-Augmented Generation (RAG), facing structural challenges from semantic and modality gaps as well as context bloat induced by multi-step retrieval. To address these challenges, we propose SYNAPSEHSG, a decoupled one-shot grounding framework for long-context multimodal Question Answering. SYNAPSEHSG introduces a hyper-semantic evidence organization method (named M2HSG) that constructs an offline graph index to align cross-modal evidence. During inference, we design a Semantic Decomposer to convert natural-language queries into graph-operable constraints. Guided by these constraints, SYNAPSEHSG retrieves a compact query-anchored subgraph in one shot and linearizes it as structured evidence for a frozen generator. This avoids repeated LLM-driven retrieval while restricting trainable components to graph construction and query parsing. To support systematic evaluation, we introduce MM-CONFQA, a new long-context multimodal QA benchmark built from academic conference presentations. Experimental results on MM-CONFQA show that SYNAPSEHSG improves grounded answer generation over strong multimodal and retrieval baselines, especially on dispersed-evidence questions, highlighting the value of pre-organized multimodal evidence over repeated search in raw heterogeneous contexts. We release the code and data to public.[1]

1 Introduction

The rapid advancement of Large Language Models (LLMs) has sparked significant interest in Multimodal Question Answering (MQA)(Alayrac et al., 2022; Liu et al., 2023). However, answering complex reasoning queries over long-context, real-world multimodal streams—such as academic conferences comprising dense PDF papers, sequential presentation slides, and continuous audio transcripts—remains challenging. Unlike standard visual question answering(Goyal et al., 2017; Marino et al., 2019), where an image and a question are statically paired, real-world MQA requires models to dynamically align and synthesize unaligned, heterogeneous evidence across extended temporal and logical spans.

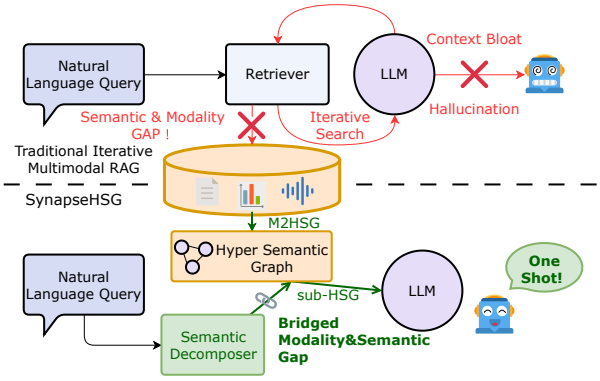

Figure 1: Traditional iterative multimodal RAG struggles with semantic and modality gaps, leading to repeated retrieval, context bloat, and hallucination. SYNAPSEHSG bridges these gaps by organizing multimodal evidence into a Hyper-Semantic Graph and retrieving a compact sub-HSG for one-shot grounded generation.

As illustrated in Figure 1, the most prevalent approaches for handling such extensive contexts is mainly with the iterative Multimodal Retrieval-Augmented Generation (RAG)(Abootorabi et al., 2025). Unfortunately, this paradigm is constrained by two core structural limitations. First, it experiences an inherent **Semantic and Modality Gap**. Standard retrievers attempt to map natural-language queries directly into a fragmented multimodal vector space. This typically results in surface-level matching(Liang et al., 2022), which

[1] https://anonymous.4open.science/r/SynapseHSG

struggles to capture the implicit logical correlations and explicit physical alignments in the multimodal stream. Second, the traditional retrieval-augmented process always need to retrieve again and again, which are so-called iterative search loops, to compensate for incomplete initial retrieval. However, it often leads to **Context Bloat** by continuously feeding the retrieved multimodal chunks back into the LLM. As the context window fills with heterogeneous and weakly aligned fragments, such as isolated slide regions without their corresponding spoken explanations, the LLM becomes more vulnerable to unsupported generation and reasoning degradation(Liu et al., 2024c; Shi et al., 2023).

To tackle these dual challenges, we propose **SYNAPSEHSG**, a decoupled, one-shot grounding framework. SYNAPSEHSG departs from iterative retrieval by combining offline hyper-semantic evidence organization with online query-conditioned subgraph grounding. This design explicitly decouples the MQA process to mitigate the modality gap and context bloat.

To bridge the modality gap, SYNAPSEHSG shifts the burden of multi-hop alignment from online retrieval to an offline structural index. We introduce **M2HSG**, which converts unaligned raw multimodal evidence into a unified **Hyper Semantic Graph** (HSG). By combining a trained association scorer with alignment-based cross-modal hyperedges constructed from precomputed multimodal correspondences, M2HSG creates a structured ontological space that explicitly aligns cross-modal elements. Simultaneously, a lightweight **Semantic Decomposer** maps the natural-language query into graph-operable structural constraints, encompassing semantic objectives, topological dependencies, and boundary conditions. This dual-sided design allows the query and evidence to meet in a shared structured semantic space.

To alleviate context bloat and hallucination risks, SYNAPSEHSG bypasses the iterative LLM search loop. Guided by the decomposed constraints above, the framework conducts multi-channel retrieval to identify a compact set of anchor nodes. It then performs deterministic hyperedge expansion to fetch physically aligned cross-modal counterparts of these anchors, isolating a well-defined, query-anchored **sub-HSG**. The result is a concise yet contextually sufficient supporting subgraph. By linearizing this subgraph(Fatemi et al., 2024), we provide a frozen generator with an explicit evidence path, enabling grounded answer generation in a single **one-shot** pass.

The main contributions of this work are threefold:

- **An Ontology-Guided Hyper-Semantic Evidence Organization Paradigm:** We introduce hyper-semantic evidence organization for long-context MQA. Implemented via M2HSG, this approach converts unaligned multimodal streams into a unified **Hyper Semantic Graph** guided by a domain-instantiable ontology. The resulting graph structurally bridges the semantic and modality gaps through high-order semantic associations and explicit cross-modal hyperedges.

- **A One-Shot Topological Grounding Mechanism:** We design a decoupled grounding architecture that systematically alleviates context bloat from iterative retrieval. A **Semantic Decomposer** translates natural-language queries into graph-operable constraints, enabling one-shot evidence selection through topological anchor expansion while keeping the heavy generator frozen.

- **A Comprehensive MQA Benchmark:** We introduce MM-CONFQA, a dedicated benchmark designed to systematically evaluate long-context multimodal QA over dense papers, sequential slides, and audio transcripts, covering diverse reasoning patterns and unanswerable queries.

Extensive evaluations on the MM-CONFQA benchmark demonstrate that SYNAPSEHSG achieves the best overall performance among the evaluated methods and yields consistent gains across different generator backbones. Further behavioral analyses show that SYNAPSEHSG improves answerable accuracy without relying on the overly conservative refusal behavior observed in traditional iterative retrieval baselines.

2 Related Work

Long-Context Multimodal Reasoning. Recent multimodal LLMs and VLMs have extended visual-language understanding from short image-text pairs to long-context multimodal inputs, including videos, documents, slides, and audio streams (Team et al., 2024; Achiam et al., 2023; Lin et al., 2024). Existing methods typically rely on extended context windows(Liu et al., 2024a), temporal token

compression(Li et al., 2024), hierarchical encoding(Carreira et al., 2022), or modality-specific pooling(Alayrac et al., 2022) to ingest larger amounts of heterogeneous input. These approaches improve input capacity and have shown strong general multimodal understanding ability. However, long-context multimodal QA often requires more than processing more tokens: relevant evidence may be distributed across visual layouts, text, and speech. Without explicit evidence organization, models may still struggle to identify which fragments should be aligned and jointly reasoned over.

Multimodal Retrieval-Augmented Generation. Retrieval-Augmented Generation (RAG) has become a common paradigm for grounding LLM outputs in external evidence (Lewis et al., 2020; Guu et al., 2020). In multimodal settings, retrieval is extended from text passages to heterogeneous units such as figures, tables, video frames, audio segments, and OCR blocks(Chen et al., 2022; Talmor et al., 2021). Existing systems often use unified embedding spaces(Radford et al., 2021; Girdhar et al., 2023), modality-specific retrievers(Chang et al., 2022), late-fusion mechanisms(Izacard and Grave, 2021), or rerankers to select evidence before generation(Nogueira and Cho, 2019). Although these methods reduce the need to process full multimodal contexts, most still treat evidence as independent chunks. This becomes limiting when the answer depends on implicit relations among fragments across modalities.

Iterative and Graph-Structured Grounding. To handle complex questions, prior work has explored iterative, multi-hop, and agentic retrieval strategies (Yao et al., 2022; Trivedi et al., 2023; Arakelyan et al., 2025), as well as graph-structured grounding methods based on entity-relation graphs, document graphs, scene graphs, event graphs, or knowledge graphs (Edge et al., 2024; Yang et al., 2025; Mo et al., 2026). These methods make evidence access more structured than flat retrieval, but iterative retrieval depends on intermediate decisions and may accumulate redundant or weakly aligned context(Asai et al., 2024; Yoran et al., 2023). Meanwhile, many graph-based methods focus on pairwise symbolic links, modality-specific structures, or localized evidence. Long-context multimodal QA requires evidence organization that can jointly account for semantic association, temporal continuity, and cross-modal alignment, motivating hyper-semantic graph construction and query-conditioned subgraph grounding.

3 Methodology

3.1 Problem Formulation and Overview

Given a natural-language question q and a multimodal evidence collection $X = \{X^{(t)}, X^{(v)}, X^{(a)}, \dots\}$, where $X^{(\mu)}$ denotes the subset of evidence units associated with modality $\mu \in \mathcal{M}$, the general multimodal question answering (MQA) task is formulated as finding the optimal answer $\hat{y}$:

$$\hat{y} = \arg\max_{y} p(y \mid q, X). \tag{1}$$

Figure 2 illustrates the proposed SYNAPSEHSG, which contains four functional stages: offline HSG construction, query decomposition, one-shot topological grounding, and grounded generation. The M2HSG first builds a hyper-semantic graph $\mathcal{G}$ from X. The Semantic Decomposer maps q into graph-operable constraints, which guide the retrieval of a compact sub-HSG $\mathcal{G}_q$. The sub-HSG is serialized as T_q and passed to a frozen LLM:

$$\hat{y} \approx \arg\max_{y} p_{\theta_{\text{LLM}}}(y \mid q, T_q). \tag{2}$$

3.2 M2HSG: Ontology-Guided Hyper-Semantic Graph Construction

M2HSG transforms the unaligned multimodal evidence X into an ontology-guided hyper-semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{H}_{evidence})$, where a domain-instantiable schema organizes semantic units and modality-grounded evidence.

Node Instantiation. M2HSG first partitions X by modality and applies a modality-specific semantic encoder ϕ_μ to extract semantic units:

$$\mathcal{V}^{(\mu)} = \phi_\mu(X^{(\mu)}), \qquad \mathcal{V} = \bigcup_{\mu \in \mathcal{M}} \mathcal{V}^{(\mu)}. \tag{3}$$

To keep graph construction query-agnostic, each extracted node $v_i \in \mathcal{V}$ is represented as a structural tuple:

$$v_i = \langle \mathbf{h}_i, z_i, \tau_i, m_i, s_i \rangle, \tag{4}$$

where $\mathbf{h}_i$ is the semantic representation, z_i is the textualized node content, τ_i is the node type, m_i is the modality, and s_i records the source location.

Pairwise Edges. To model semantic correlations, a trainable association scorer f_{θ_g} predicts the retention probability of each candidate pair:

$$\hat{p}_{ij} = \sigma\big(f_{\theta_g}(\mathbf{u}_{ij})\big), \tag{5}$$

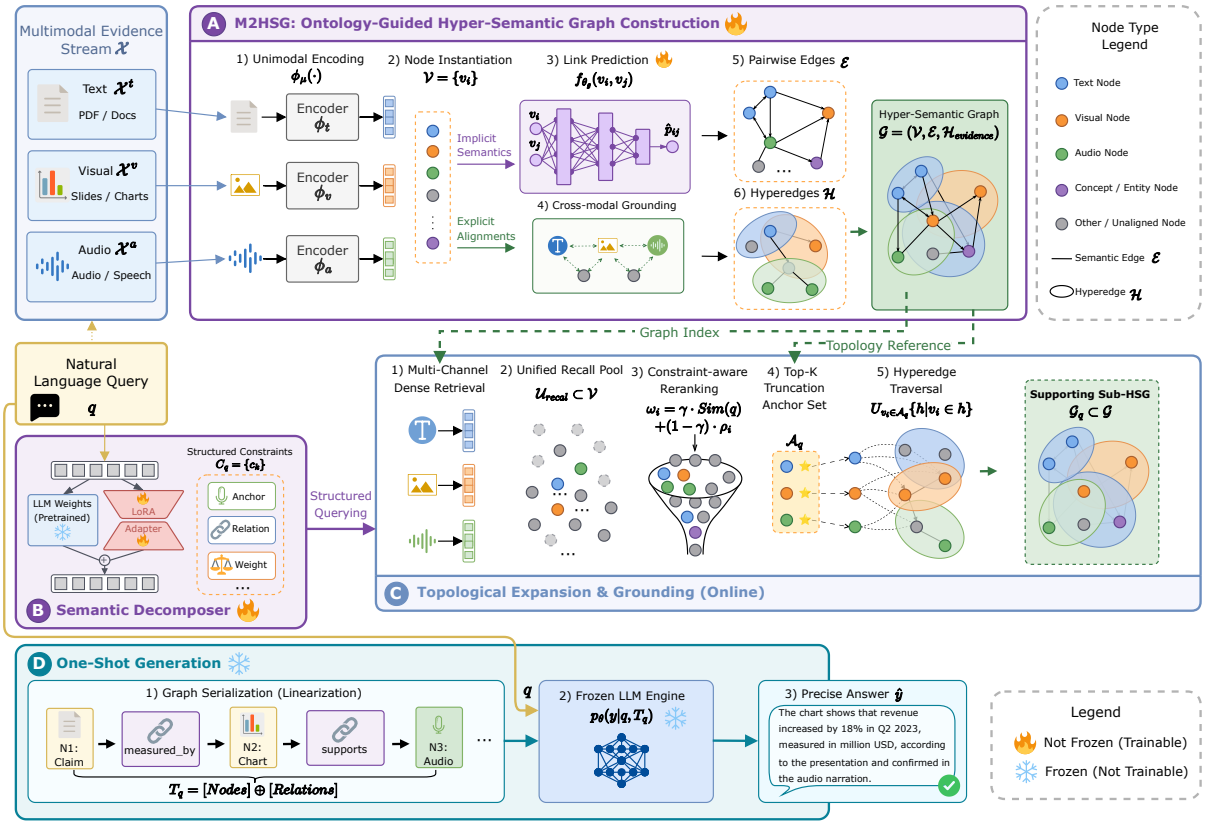

Figure 2: Overall architecture of SYNAPSEHSG. The framework contains four functional stages: ontology-guided HSG construction with M2HSG, query-aware constraint generation with the Semantic Decomposer, one-shot topological grounding over the graph, and grounded answer generation with a frozen LLM. Trainable and frozen components are marked separately.

where $\mathbf{u}_{ij}$ concatenates the semantic representations, node-type embeddings, and modality embeddings of v_i and v_j. The pairwise edge set is obtained by thresholding:

$$\mathcal{E} = \{(v_i, v_j) \mid \hat{p}_{ij} > \delta_e\}. \tag{6}$$

The optimization objective for the graph scoring module is given in Section 3.5.

Cross-modal Hyperedges. Beyond binary pairwise edges, multimodal evidence often forms high-order cross-stream correspondences. M2HSG first constructs candidate cross-modal hyperedges through deterministic alignment and retrieval cues. Each candidate hyperedge groups a semantic center node with its aligned evidence units across modalities:

$$h_k = \{v_{k_1}, v_{k_2}, \ldots, v_{k_{r_k}}\}, \quad h_k \in \mathcal{H}_{evidence}. \tag{7}$$

To calibrate these candidate structures, we assign each hyperedge a coherence score:

$$\psi(h_k) = g_{\theta_g}\left(\left\{\mathbf{h}_{k_t}, \mathbf{e}_{\tau_{k_t}}, \mathbf{e}_{m_{k_t}}\right\}_{t=1}^{r_k}\right), \tag{8}$$

where g_{θ_g} measures cross-modal consistency. Retained hyperedges serve as evidence links for deterministic anchor expansion during inference.

3.3 Semantic Decomposer and Multi-Channel Retrieval

While the hyper-semantic graph $\mathcal{G}$ provides a structured evidence space, the input question q resides in a unstructured natural language space. To align the question with the hyper-semantic graph, SYNAPSEHSG introduces a neural Semantic Decomposer D_{θ_d} that maps q to a set of graph-operable constraints:

$$\mathcal{C}_q = D_{\theta_d}(q) = \{c_k\}_{k=1}^{K}. \tag{9}$$

Formally, each decomposed constraint is instantiated as a structural tuple:

$$c_k = \langle a_k, r_k, w_k \rangle, \tag{10}$$

where a_k specifies the semantic query anchor, r_k defines the expected ontology-compatible relation within $\mathcal{E}$, and w_k denotes its importance weight. The optimization regime is detailed in Section 3.5.

Conditioned on $\mathcal{C}_q$, SYNAPSEHSG performs parallel multi-channel dense retrieval using the base question q and each semantic anchor a_k as independent retrieval queries. Their top-ranked results are aggregated into a unified recall pool:

$$\mathcal{U}_{recall} = \mathcal{R}_q \cup \bigcup_{k=1}^{K} \mathcal{R}_k \subset \mathcal{V}, \tag{11}$$

$$\mathcal{R}_x = \mathrm{TopK}_{B_x}\left(\{v_i \in \mathcal{V}\}, \mathrm{sim}(\mathbf{h}_i, \mathbf{x})\right), \tag{12}$$

where $\mathbf{x}$ denotes the dense encoding of retrieval query x, $x \in \{q, a_1, \ldots, a_K\}$. Retrieval budgets are detailed in Appendix F.

3.4 Topological Expansion and Grounding

To obtain a compact query-conditioned subgraph, SYNAPSEHSG executes constraint-aware reranking and hyperedge traversal.

Constraint-Aware Reranking. For each candidate $v_i \in \mathcal{U}_{recall}$, the relevance score ω_i is formulated as the linear combination of the dense query similarity and a constraint compatibility score ρ_i:

$$\omega_i = \gamma \cdot \mathrm{sim}(\mathbf{h}_i, \mathbf{q}) + (1-\gamma) \cdot \rho_i. \tag{13}$$

The compatibility score explicitly incorporates the decomposed anchors, topological relations, and importance weights:

$$\rho_i = \sum_{k=1}^{K} w_k \Big[\alpha \cdot \mathrm{sim}(\mathbf{h}_i, \mathbf{a}_k) + (1-\alpha) \cdot \eta(v_i, r_k, \mathcal{G})\Big] + \chi(\tau_i), \tag{14}$$

where $\eta(v_i, r_k, \mathcal{G})$ measures whether candidate node v_i satisfies the expected topological relation r_k in $\mathcal{G}$, and $\chi(\tau_i)$ denotes a node-type compatibility bonus. The recall pool is then truncated to an anchor set $\mathcal{A}_q$ bounded by an anchor budget B_a:

$$\mathcal{A}_q = \mathrm{TopK}\left(\mathcal{U}_{recall}, \omega(\cdot), B_a\right), \tag{15}$$

where $\omega(v_i) = \omega_i$. The selected anchors are used as entry points for the subsequent hyperedge traversal and sub-HSG construction.

Hyperedge Traversal and Linearization. To capture comprehensive multimodal evidence, SYNAPSEHSG expands the anchor set into a grounded node set $\mathcal{V}_q$ via hyperedge traversal:

$$\mathcal{V}_q = \mathcal{A}_q \cup \{u \mid \exists v_i \in \mathcal{A}_q, \tag{16}$$

where $h \in \mathcal{H}_{evidence} : \{u, v_i\} \subseteq h\}$. The retained query-relevant hyperedges are defined as:

$$\mathcal{H}_q = \{h \in \mathcal{H}_{evidence} \mid h \cap \mathcal{A}_q \neq \emptyset\}. \tag{17}$$

The induced pairwise edge set is:

$$\mathcal{E}_q = \{(v_i, v_j) \in \mathcal{E} \mid v_i, v_j \in \mathcal{V}_q\}. \tag{18}$$

The induced sub-HSG is then defined as $\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q, \mathcal{H}_q)$. For generation, SYNAPSEHSG serializes the grounded nodes and induced pairwise relations into $T_q = [T_{\mathcal{V}} \oplus T_{\mathcal{E}}]$, where nodes retain their identifiers, modalities, types, source locations, and textualized contents, while relations retain their source nodes, relation types, and target nodes. The serialized context T_q is concatenated with q as the input to the frozen LLM.

3.5 Training Strategy

SYNAPSEHSG adopts a decoupled optimization regime. The modality encoders and final LLM generator are frozen, while the trainable parameters are restricted to the graph scorer θ_g and the query decomposer θ_d.

Optimization of M2HSG. The graph scoring module is trained with two objectives: pairwise edge retention and hyperedge-level cross-modal coherence. For a candidate pair (v_i, v_j) with edge label $y_{ij} \in \{0, 1\}$, the edge objective is:

$$\mathcal{L}_{\text{edge}} = -\sum_{(v_i, v_j)} \Big[y_{ij} \log \hat{p}_{ij} + (1 - y_{ij}) \log(1 - \hat{p}_{ij})\Big], \tag{19}$$

where $\hat{p}_{ij}$ is the predicted retention probability. For a candidate hyperedge h_k with coverage label $y_k^h \in \{0, 1\}$, the hyperedge objective is:

$$\mathcal{L}_{\text{hyper}} = -\sum_{h_k} \Big[y_k^h \log \psi(h_k) + (1 - y_k^h) \log(1 - \psi(h_k))\Big], \tag{20}$$

where $y_k^h = 1$ indicates that the candidate hyperedge covers evidence from at least two modalities. The final graph construction loss is:

$$\mathcal{L}_{\text{graph}} = \mathcal{L}_{\text{edge}} + \mathcal{L}_{\text{hyper}}. \tag{21}$$

Both objectives are optimized within the same graph scoring module parameterized by θ_g.

Optimization of the Semantic Decomposer. The query decomposer D_{θ_d} is trained with LoRA-based(Hu et al., 2022) supervised fine-tuning to generate structural constraints:

$$\mathcal{L}_{\text{decomp}} = -\sum_{k=1}^{|\mathcal{C}_q|} \log p_{\theta_d}(c_k^* \mid c_{<k}^*, q), \tag{22}$$

where c_k^* denotes the target constraint sequence constructed from training QA annotations and source-level evidence metadata, as detailed in Appendix E.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate SYNAPSEHSG on **MM-ConfQA**, a multimodal question-answering dataset constructed from 1,553 academic presentations at ICLR and NeurIPS (2024-2025). The dataset comprises 11,873 QA pairs, where the validation (400 samples) and test (500 samples) splits are human-verified gold standards, while the remaining 10,973 pairs form the training set. These queries are grounded in three aligned modalities: paper text (PDF), slide images (PNG), and audio tracks (MP3). Statistically, a large portion of answerable queries require evidence from at least two modalities, while 12% are unanswerable to evaluate evidence calibration. Details regarding the construction pipeline and filtering heuristics are deferred to Appendix A.

Evaluation Metrics. Since answers are often free-form and may admit multiple semantically equivalent realizations, we evaluate model predictions using an LLM-based semantic matching protocol. Specifically, GPT-5.4 is used as the evaluator to determine whether a prediction is equivalent to the reference answer under a fixed judging rubric(Zheng et al., 2023). To avoid conflating answer generation with abstention behavior(Rajpurkar et al., 2018), we report three complementary metrics. **Overall Accuracy (OA)** measures correctness over the full test set. **Answerable Accuracy (Ans. Acc.)** is computed only on questions with sufficient supporting evidence. **Unanswerable Refusal Accuracy (Unans. Ref.)** measures the proportion of unanswerable questions for which the model correctly abstains instead of producing an unsupported answer.

Baselines. To ensure a comprehensive evaluation, we benchmark the proposed SYNAPSEHSG against several strong baselines across three prevailing paradigms. We first evaluate against several **Multimodal RAG Methods** designed for cross-modal dense retrieval, such as VisRAG (Yu et al., 2025), RAG-Anything (Guo et al., 2025), and M2RAG (Ma et al., 2025). As structural alternatives to our holistic hyperedge modeling, we further compare with **Iterative RAG Methods**, including pioneering works like EventRAG (Yang et al., 2025), KGGen (Mo et al., 2026), and GraphRAG (Edge et al., 2024), which rely on multi-step graph exploration. Additionally, to assess native reasoning capabilities without explicit retrieval modules, we compare SYNAPSEHSG with **Direct Multimodal LLMs**, encompassing representative open-source and proprietary models such as LLaVA-1.6-7B (Liu et al., 2024b), Qwen3.6-27B (Qwen Team, 2026b), and GPT-5.4 (OpenAI, 2026). For a strictly controlled comparison, all applicable retrieval-based and iterative baselines use **Qwen3.6-27B** as the generator. Finally, to evaluate backbone robustness, we instantiate SYNAPSEHSG with LLaVA-1.6-7B, Qwen3.6-27B, and GPT-5.4.

4.2 Main Results

Table 1 summarizes the main results on the MM-ConfQA benchmark, and Figures 3, 4, and 5 provide further analyses of generator robustness, answer-refusal behavior, and reasoning-pattern performance. SYNAPSEHSG achieves the highest overall accuracy among the evaluated methods. We highlight four observations below.

Effectiveness under controlled settings. When using the same Qwen3.6-27B generator, SYNAPSEHSG achieves 72.2% overall accuracy and 73.4% answerable accuracy. It improves over the strongest multimodal RAG baseline, M2RAG, by 7.0 percentage points in overall accuracy, and over the graph-based baseline, GraphRAG, by 13.8 percentage points. This margin suggests that generic multimodal retrieval and iterative graph-based evidence exploration do not fully address MM-ConfQA, where relevant cues are often dispersed across modalities and slides. In contrast, query-conditioned organization of hyper-semantic evidence provides a more effective context for grounded answer generation.

Robustness across generator backbones. As shown in Figure 3, SYNAPSEHSG consistently improves different generators. It increases overall accuracy by 7.8, 7.6, and 8.4 points for LLaVA-1.6-7B, Qwen3.6-27B, and GPT-5.4, respectively. This trend indicates that the benefit of SYNAPSEHSG is not tied to a specific backbone, but comes from the structured organization and query-conditioned selection of supporting evidence.

Answerability vs. refusal trade-off. Figure 4 plots answerable accuracy against unanswerable re-

Method	Overall Performance			Reasoning Pattern Accuracy					
	Overall	Ans. Acc.	Unans. Ref.	Single	Link	Multi	Cross	Comp.	Caveat
Direct Multimodal LLMs (Zero-shot baselines)									
LLaVA-1.6-7B	45.6	45.9	43.3	32.5	40.0	38.7	53.3	63.3	55.0
Qwen3.6-27B	64.6	66.8	48.3	52.5	63.3	57.3	76.0	80.0	75.0
GPT-5.4	68.8	67.0	81.7	55.0	63.3	68.0	74.7	70.0	72.5
Multimodal RAG Methods (Generator: Qwen3.6-27B)									
VisRAG	57.4	56.1	66.7	22.5	50.7	52.0	70.7	68.3	72.5
RAG-Anything	57.8	60.2	40.0	47.5	55.3	52.0	68.0	70.0	77.5
M2RAG	64.2	66.4	48.3	47.5	62.7	53.3	73.3	86.7	80.0
Iterative RAG Methods (Generator: Qwen3.6-27B)									
EventRAG	43.0	35.9	**95.0**	37.5	29.3	37.3	37.3	46.7	37.5
KGGen	50.2	45.0	88.3	37.5	40.7	36.0	50.7	56.7	57.5
GraphRAG	59.6	60.2	55.0	32.5	52.0	62.7	72.0	71.7	75.0
SynapseHSG *(Ours)*									
SynapseHSG (w/ LLaVA-1.6-7B)	53.4	52.7	58.3	45.0	52.0	46.7	60.0	55.0	57.5
SynapseHSG (w/ Qwen3.6-27B)	72.2	73.4	63.3	**62.5**	65.3	69.3	82.7	85.0	87.5
SynapseHSG (w/ GPT-5.4)	**77.2**	**80.9**	50.0	60.0	**78.0**	**81.3**	**85.3**	**90.0**	**90.0**

Table 1: Main quantitative results on the MM-ConfQA benchmark. *Overall* denotes accuracy over the entire test set, *Ans. Acc.* measures accuracy exclusively on answerable queries, and *Refusal* quantifies the correct abstention rate on unanswerable questions. The best results are highlighted in **bold**, and the second-best results are underlined.

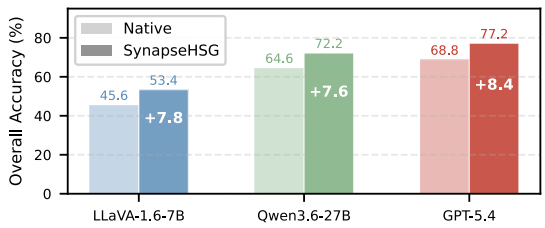

Figure 3: Generator robustness analysis. SynapseHSG consistently improves overall accuracy across LLaVA-1.6-7B, Qwen3.6-27B, and GPT-5.4, with gains of 7.8, 7.6, and 8.4 points, respectively.

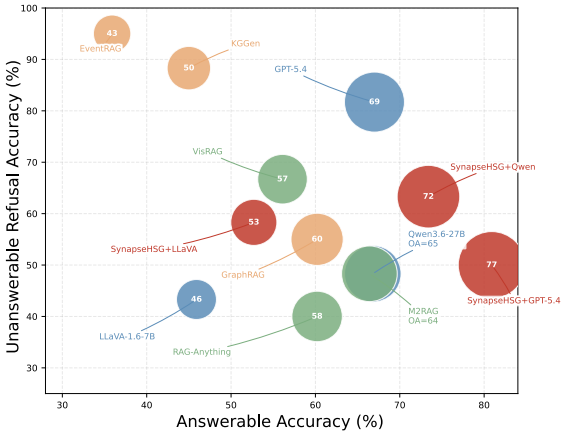

Figure 4: Answer-refusal trade-off. The x-axis reports answerable accuracy, the y-axis reports correct refusal accuracy on unanswerable questions, and the bubble size and number indicate overall accuracy. Some methods achieve high refusal accuracy at the cost of lower answerable accuracy, while SynapseHSG improves overall performance primarily through stronger answerable-question accuracy, reflecting a different answerability–abstention trade-off.

fusal accuracy, where the bubble numbers and sizes represent the overall accuracy (OA) of each configuration. EventRAG and KGGen achieve high refusal accuracy, reaching 95.0% and 88.3%, respectively, but obtain much lower answerable accuracy, with 35.9% and 45.0%. This visualization indicates that high refusal accuracy alone is not sufficient to demonstrate robust grounded QA, since it may also reflect overly conservative behavior. In contrast, SynapseHSG improves overall accuracy mainly through stronger performance on answerable questions, achieving 73.4% answerable accuracy with Qwen3.6-27B and 80.9% with GPT-5.4. Its refusal accuracy is moderate rather than maximal, reflecting a different trade-off from methods with high refusal accuracy but weak answerable-question performance.

Improvements on complex reasoning patterns. The reasoning-pattern breakdown in Table 1 and the heatmap in Figure 5 show that SynapseHSG mainly improves evidence-grounded reasoning categories. With Qwen3.6-27B, it substantially improves over the native generator across categories, with the largest gains on *Multi-Hop Aggregation* (+12.0%) and *Limitation/Caveat* (+12.5%) questions. This suggests that the graph structure is most useful when dispersed evidence must be connected before generation.

The same trend appears with GPT-5.4. SynapseHSG improves over the native generator on *Evidence Linking* (63.3% to 78.0%), *Multi-Hop*

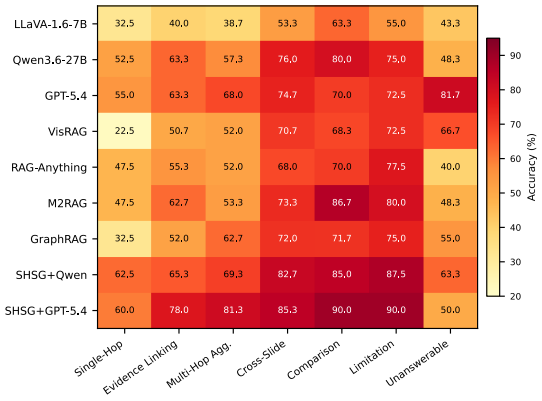

Figure 5: Reasoning-pattern breakdown on MM-ConfQA. SHSG denotes SynapseHSG. The results show that SynapseHSG yields larger gains on evidence-intensive categories, including multi-hop aggregation, cross-slide reasoning, and limitation/caveat questions.

Agg. (68.0% to 81.3%), *Cross-Slide* reasoning (74.7% to 85.3%), *Comparison* (70.0% to 90.0%), and *Limitation/Caveat* questions (72.5% to 90.0%). These results indicate that stronger generators still benefit from structured evidence organization. Meanwhile, the gains are not uniform across all categories; for instance, the accuracy on *Unanswerable* queries displays varying trends across backbones. This suggests that the main advantage of SynapseHSG lies in grounded answer generation over dispersed multimodal evidence.

Overall, the empirical results indicate that organizing context into query-anchored hyper-semantic structures provides consistent gains across different generator backbones and evidence-intensive reasoning patterns, supporting its effectiveness for conference-oriented multimodal QA.

4.3 Ablation Study

To examine component contributions, we ablate SynapseHSG with Qwen3.6-27B as the generator. Figure 6 reports changes in overall accuracy, answerable accuracy, and unanswerable refusal accuracy relative to the full model.

Removing the hyper-semantic graph causes the largest degradation. The w/o HSG variant drops by 9.0 points in overall accuracy and 9.3 points in answerable accuracy, indicating that the structured evidence space is central to grounded answer generation. Replacing the trainable HSG construction with heuristic edges also reduces performance (-2.4 overall, -2.9 answerable), suggesting that learned association scoring provides more useful evidence organization than heuristic linking alone.

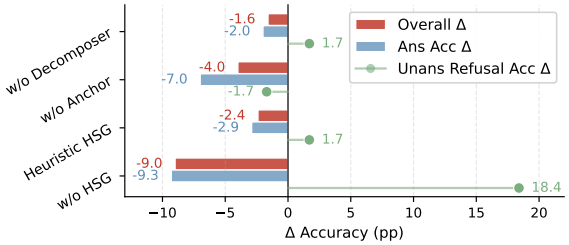

Figure 6: Ablation study of SynapseHSG. Changes are computed relative to the full model. Removing the hyper-semantic graph causes the largest drop, while removing anchors reduces answerable accuracy and increases refusal accuracy.

The anchor mechanism also contributes substantially. Removing anchors reduces overall and answerable accuracy by 4.0 and 7.0 points, respectively, while increasing unanswerable refusal accuracy by 18.4 points. This shows that the variant becomes more conservative rather than more accurate on answerable questions, confirming the need to report answerable and unanswerable metrics separately.

Removing the decomposer causes a smaller but consistent decline (-1.6 overall, -2.0 answerable). Overall, the ablation results show that the hyper-semantic graph provides the largest gain, while anchor-based grounding and query decomposition further improve answerable-question performance.

5 Conclusion

We introduced SynapseHSG, a decoupled one-shot grounding framework for long-context multimodal QA. By organizing multimodal evidence into a Hyper-Semantic Graph with M2HSG and retrieving compact query-anchored subgraphs through semantic decomposition, SynapseHSG reduces reliance on iterative retrieval while addressing semantic and modality gaps. We also introduced MM-ConfQA, a benchmark for conference-oriented multimodal QA. Experiments show that SynapseHSG outperforms strong multimodal and retrieval-based baselines, with larger gains on evidence-intensive reasoning patterns. These results support pre-aligned structural grounding as an effective substrate for long-context multimodal reasoning.

Limitations

SYNAPSEHSG has several limitations that motivate future research. Although the framework is designed to be modality-agnostic, its ontology requires domain-specific instantiation. This work instantiates the ontology for academic presentations; applying the framework to medical records, financial reports, meetings, lectures, or open-domain videos requires adapting the node taxonomy, relation schema, and upstream preprocessing pipeline.

As a pipelined architecture, SYNAPSEHSG is subject to cascading errors. Imperfections in multimodal parsing, temporal alignment, or structural encoding during offline graph construction may propagate to the retrieved subgraphs and affect final generation. The topological density and granularity of the constructed HSG are also sensitive to extraction heuristics, filtering thresholds, and graph-scoring configurations.

The framework relies on the Semantic Decomposer to map natural-language queries into ontology-compatible constraints. Ambiguous or out-of-distribution queries may yield underspecified structural constraints, which can hinder anchor selection and cross-modal hyperedge expansion.

Final generation quality remains influenced by the capacity of the frozen LLM. If the retrieved sub-HSG omits relevant evidence or the generator fails to synthesize the multi-hop context faithfully, the system may produce incomplete or unsupported answers. Developing faithfulness verification mechanisms over serialized sub-HSGs remains an important direction for future work.

Ethics Statement

This work introduces MM-CONFQA, constructed from publicly accessible academic presentations. To respect intellectual-property constraints and platform terms of service, we release source identifiers and benchmark annotations, but do not redistribute raw papers, slide images, or audio recordings. The source recordings may contain personally identifiable information already present in public conference materials, such as speaker voices, names, and affiliations. We do not introduce additional demographic profiling or biographical annotations. The authors declare no conflicts of interest.

SYNAPSEHSG relies on automatic extraction, retrieval, and LLM-based generation. It may generate unsupported claims when cross-modal evidence is incomplete or misretrieved, so outputs should be verified against the original materials in high-stakes applications. Our LLM-as-a-judge evaluation protocol uses a standardized rubric across methods, but automated evaluation may introduce bias or inconsistency. Computationally, the framework keeps the final generator frozen and restricts trainable parameters to the graph scorer and Semantic Decomposer, reducing the need for full-parameter multimodal fine-tuning, although offline preprocessing and graph construction still require computational resources.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Erik Arakelyan, Pasquale Minervini, Patrick Lewis, Pat Verga, and Isabelle Augenstein. 2025. Flare: Faithful logic-aided reasoning and exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23396–23414.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pages 9112–9141.

Joao Carreira, Skanda Koppula, Daniel Zoran, Adria Recasens, Catalin Ionescu, Olivier Henaff, Evan Shelhamer, Relja Arandjelovic, Matt Botvinick, Oriol Vinyals, and 1 others. 2022. Hip: Hierarchical perceiver. *arXiv preprint arXiv:2202.10890*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16495–16504.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *Preprint*, arXiv:2402.03216.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. 2024. Talk like a graph: Encoding graphs for large language models. In *International conference on learning representations*, volume 2024, pages 43909–43934.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 5971–5984.

Hao Liu, Matei Zaharia, and Pieter Abbeel. 2024a. Ringattention with blockwise transformers for near-infinite context. In *International Conference on Learning Representations*, volume 2024, pages 3992–4008.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024c. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Zi-Ao Ma, Tian Lan, Rong-Cheng Tu, Yong Hu, Yu-Shi Zhu, Tong Zhang, Heyan Huang, and Xian-Ling Mao. 2025. Multi-modal retrieval augmented multi-modal generation: Datasets, evaluation metrics and strong baselines. *Preprint*, arXiv:2411.16365.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Belinda Mo, Kyssen Yu, Joshua Kazdan, Proud Mpala, Lisa Yu, Charilaos Kanatsoulis, and Sanmi Koyejo. 2026. Kggen: Extracting knowledge graphs from plain text with language models. *Advances in Neural Information Processing Systems*, 38:30092–30115.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

OpenAI. 2026. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/. Accessed: 2026-05-18.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Qwen Team. 2026a. Qwen3.5: Towards native multimodal agents.

Qwen Team. 2026b. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

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

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 10014–10037.

Zairun Yang, Yilin Wang, Zhengyan Shi, Yuan Yao, Lei Liang, Keyan Ding, Emine Yilmaz, Huajun Chen, and Qiang Zhang. 2025. EventRAG: Enhancing LLM generation with event knowledge graphs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16967–16979, Vienna, Austria. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2025. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. In *International Conference on Learning Representations*, volume 2025, pages 21074–21098.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A MM-ConfQA Dataset Details

MM-ConfQA is constructed from 1,553 academic conference presentations collected from SlidesLive[2], covering ICLR[3] and NeurIPS[4] in 2024–2025, as detailed in Table 2. Each presentation is associated with three modalities from the original source: a paper PDF, a slide deck, and an audio recording.

The benchmark comprises 11,873 QA pairs, partitioned into training, validation, and test splits (Table 3). The validation and test splits are human-verified by domain researchers and used as gold-standard evaluation sets. The questions cover diverse reasoning patterns, including single-hop retrieval, evidence linking, multi-hop aggregation, cross-slide reasoning, comparison, limitation/caveat identification, and unanswerable cases. Table 4 summarizes the test-set distribution of these reasoning patterns.

Release Strategy. To respect the intellectual-property constraints of the original conference materials and comply with platform terms of service, we do not directly redistribute the raw PDFs, slide images, or audio files. Instead, the release provides SlidesLive `session_ids` and benchmark annotations, including QA pairs, answerability labels, answer types, reasoning-pattern metadata, and split information. Researchers can use these identifiers to access the corresponding source materials through the official platform. During evaluation, no oracle evidence paths are provided; systems are expected to retrieve or construct supporting evidence from the available multimodal sources.

Venue	Year	Sessions
ICLR	2024	226
NeurIPS	2024	215
ICLR	2025	631
NeurIPS	2025	481
Total	–	**1,553**

Table 2: Source distribution of MM-ConfQA.

Split	Samples	Answerable	Unanswerable
Train	10,973	9,818	1,155
Validation	400	352	48
Test	500	440	60
Total	**11,873**	**10,610**	**1,263**

Table 3: QA split statistics of MM-ConfQA.

Reasoning Pattern	Count	Description
Evidence Linking	150	Trace a claim through supporting evidence
Multi-Hop Aggregation	75	Synthesize multiple evidence pieces
Cross-Slide	75	Use evidence distributed across slides
Comparison	60	Compare methods, results, or settings
Unanswerable	60	No valid answer exists in the evidence
Single-Hop	40	Retrieve one direct evidence unit
Limitation/Caveat	40	Identify limitations or caveats

Table 4: Reasoning-pattern distribution of the MM-ConfQA test set.

B Implementation Details

We implement SynapseHSG as a decoupled offline-online pipeline. Algorithm 1 summarizes the full execution flow. The offline stage constructs the M2HSG index from paper text, slide regions, transcript utterances, and audio-derived cues. The online stage performs query decomposition, multi-channel retrieval, constraint-aware reranking, hyperedge anchor expansion, structure-preserving linearization, and frozen-LLM generation.

The final answer generator is kept frozen across all experiments. The trainable components are restricted to the graph scoring module and the Semantic Decomposer. Detailed supervision and hyperparameter settings for both modules are provided in Appendix D and Appendix E.

For multimodal preprocessing, we use `Whisper-large-v3`(Radford et al., 2022) for audio transcription and timestamped speech segments. Evidence chunks are encoded with `BGE-M3`(Chen et al., 2024), and `FAISS`(Douze et al., 2024) is used for session-scoped vector search over HSG nodes. Local training, preprocessing, embedding, retrieval, and open-weight generator inference are conducted on 4× NVIDIA A100 80G GPUs. Qwen3.6-27B and LLaVA-1.6-7B are deployed locally with vLLM(Kwon et al., 2023), while GPT-5.4 is accessed through its API. We use random seed 42 for trainable components.

[2] https://slideslive.com/
[3] https://iclr.cc/
[4] https://neurips.cc/

Algorithm 1 SYNAPSEHSG Offline Construction and Online Grounding

Require: Multimodal evidence collection X, question q
Ensure: Answer $\hat{y}$
Offline Construction
Extract ontology-guided nodes $\mathcal{V}$ from paper, slide, and audio evidence
Generate candidate pairwise relations and score them with the graph scorer
Construct candidate cross-modal hyperedges from alignment and retrieval cues
Retain pairwise edges $\mathcal{E}$ and evidence hyperedges $\mathcal{H}_{evidence}$
Build the FAISS node index and node-to-triple inverted index
Online Grounding
Generate structural constraints $\mathcal{C}_q = D_{\theta_d}(q)$ with the Semantic Decomposer
Query the FAISS node index with the base question and decomposed constraints
Map retrieved nodes to candidate triples through the inverted index
Merge candidates into $\mathcal{U}_{recall}$ and rerank them with constraint compatibility
Select top-ranked seed triples and perform hyperedge anchor expansion
Construct the grounded sub-HSG $\mathcal{G}_q = (\mathcal{V}_q, \mathcal{E}_q, \mathcal{H}_q)$
Linearize $\mathcal{G}_q$ into $T_q = [T_{\mathcal{V}} \oplus T_{\mathcal{E}}]$
Generate $\hat{y}$ with the frozen LLM conditioned on (q, T_q)
return $\hat{y}$

C Ontology Instantiation for MM-CONFQA

SYNAPSEHSG adopts an ontology-guided graph construction mechanism. The ontology provides a domain-instantiable schema for organizing semantic units, modality-grounded evidence, and typed cross-modal associations. For MM-CONFQA, we instantiate this schema for academic conference presentations.

Table 5 lists the node taxonomy used in this instantiation. The schema separates *Semantic Nodes*, which represent abstract research concepts, from *Evidence-Grounded Nodes*, which refer to localized modality-specific artifacts. Table 6 lists the relation taxonomy used to connect these nodes. Table 7 summarizes the type constraints used when constructing candidate relations.

Node Type	Description
Semantic Nodes	
Claim	Core scientific claim or contribution
Method	Methodological component or algorithmic design
Result	Experimental finding or reported outcome
Metric	Evaluation metric or measurement criterion
Limitation	Stated limitation, caveat, or failure case
Evidence-Grounded Nodes	
PaperChunk	Textual evidence segment from the paper
Slide	Slide-level presentation unit
VisualRegion	Local visual or layout region from a slide or figure
Utterance	Transcript span from the presentation audio
ProsodyEvent	Audio emphasis cue, such as acoustic stress or pause
Figure/Table	Figure or table reference from paper or slide material

Table 5: Node taxonomy used in the academic-presentation ontology instantiated for MM-CONFQA.

Relation Type	Description
Semantic Relations	
`supported_by`	Links a claim or result to supporting evidence
`measured_by`	Links a result to an evaluation metric
`compares`	Links methods or results involved in a comparison
`referenced_by`	Captures cross-references among evidence nodes
Cross-Modal Grounding Relations	
`grounded_in_paper`	Grounds a node in paper text, figures, or tables
`aligned_to_slide`	Aligns semantic or spoken content with slide layouts
`emphasizes`	Links audio prosody cues to emphasized semantic content

Table 6: Relation taxonomy used in the academic-presentation ontology instantiated for MM-CONFQA.

D Graph Scorer Supervision and Training

The graph scorer is trained to filter candidate pairwise relations and calibrate hyperedge-level coherence during offline M2HSG construction. Since exhaustive human annotation of dense multimodal graph edges is difficult to scale to the full training split, we construct auxiliary structural supervision from training-set QA annotations and source-level evidence metadata. These automatically derived labels are used only to optimize the offline M2HSG construction module and are not provided as oracle evidence during evaluation.

Auxiliary Supervision. For the pairwise edge-retention objective, a candidate edge (v_i, v_j) is assigned a positive label ($y_{\text{edge}} = 1$) if its incident nodes belong to high-confidence evidence paths

Relation Type	Allowed Head Node Types	Allowed Tail Node Types
supported_by	Claim, Result	PaperChunk, VisualRegion, Figure/Table, Result
measured_by	Result	Metric
compares	Method, Result	Method, Result
grounded_in_paper	Claim, Method, Result, Slide, VisualRegion	PaperChunk, Figure/Table
aligned_to_slide	Utterance, Claim, Method, Result	Slide, VisualRegion
emphasizes	ProsodyEvent	Claim, Method, Result, Metric
referenced_by	Figure/Table, VisualRegion, PaperChunk	Utterance, PaperChunk, Claim, Method

Table 7: Type constraints used by M2HSG when constructing candidate relations in the MM-ConfQA ontology instantiation.

that support the reference answer. The matching protocol evaluates semantic overlap among the node content, decomposed question intent, and reference answer under source-level provenance constraints. Candidate edges that fail to satisfy these criteria are sampled as negative instances ($y_{\text{edge}} = 0$).

For the hyperedge-coherence objective, a candidate hyperedge h_k receives a positive coverage label ($y_{\text{hyper}} = 1$) if its linked evidence nodes span at least two distinct modalities. This objective encourages the M2HSG scorer to retain cross-modal evidence links rather than relying on isolated unimodal artifacts.

Scorer Implementation. The scorer is implemented as a two-hidden-layer MLP with hidden dimension 128, GELU activation, and dropout 0.1 after the first hidden layer. It operates over feature vectors derived from candidate relation metadata, semantic similarity signals, modality indicators, and relation-type embeddings. For each candidate relation, we construct sparse lexical and structural features, including text overlap, dense semantic similarity, relation identity, node-type indicators, and modality-pair indicators. To control the high-dimensional sparse feature space and improve training efficiency, we cap raw sparse features at 5,000 dimensions and project them into a 512-dimensional dense representation using TruncatedSVD(Pedregosa et al., 2011). A trainable 16-dimensional relation embedding is concatenated with the compressed feature vector before being passed to the MLP scorer.

The scorer is trained on a triple-level split after filtering candidates with `include_in_loss=true`, resulting in approximately 222K training instances. It is optimized with AdamW(Loshchilov and Hutter, 2017) using a learning rate of 1×10^{-3}, batch size 64, weight decay 1×10^{-4}, and 25 training epochs. We apply relation-balanced sample weights and select the best checkpoint by maximizing the sum of Edge AP and Hyperedge AP on the development split. The training objective follows the pairwise edge-retention and hyperedge-coherence losses described in Section 3.5.

Feature Group	Description	Dimension
Raw High-Dimensional Features		
Lexical Overlap	TF-IDF, BM25, string matching	up to 2,500
Semantic Affinity	Dense embedding similarity	up to 2,000
Modality Indicators	Source modality and modality-pair masks	up to 500
Dense Latent Representations		
SVD-Compressed Vector	TruncatedSVD projection	512
Relation Embedding	Trainable relation-type embedding	16

Table 8: Feature groups used by the M2HSG graph scorer. Raw sparse features are capped at 5,000 dimensions and projected into a compact dense representation before being passed to the MLP scorer.

E Semantic Decomposer Supervision

The Semantic Decomposer maps a natural-language question into ontology-compatible structural constraints. Each output follows the JSON schema summarized in Table 9. The decomposer is initialized from `Qwen3.5-9B`(Qwen Team, 2026a) and trained with LoRA-based supervised fine-tuning on 10,973 training questions, with 1,316 held-out decomposition examples used for checkpoint selection.

The training prompt requires valid JSON without markdown or explanatory text. The generated fields are used as supervision targets for query in-

Field	Type	Role	Allowed Values / Structure
`target_semantic_unit`	Enum	Semantic node type used to localize the query focus.	`Claim`, `Method`, `Result`, `Metric`, `Limitation`
`target_relation`	Enum	Query-side relation label used as a structural intent signal.	`supported_by`, `emphasizes`, `compares`, `measured_by`, `aligned_to_slide`, `grounded_in_paper`, `other`
`constraints`	Object	Structured filters for retrieval channels and reranking compatibility.	`time_constraint`, `page_constraint`, `baseline_constraint`, `metric_constraint`, `section_constraint`, `experiment_setting`
`modality_requirement`	Enum	Expected evidence modality or modality combination.	`paper`, `audio`, `visual`, `paper+audio`, `paper+visual`, `audio+visual`, `paper+audio+visual`, `none`
`reasoning_pattern`	Enum	Reasoning category predicted from the question.	`single_hop`, `evidence_linking`, `cross_slide`, `cross_time`, `comparison`, `limitation_or_caveat`, `multi_hop_aggregation`, `unanswerable`
`answer_type`	Enum	Expected answer format.	`entity`, `phrase`, `number`, `explanation`, `yes_no`, `abstain`
`answerable`	Boolean	Whether sufficient evidence is expected to exist.	`true`, `false`

Table 9: Output schema and allowed values of the Semantic Decomposer. Relation labels correspond to the effective query-side inventory covered by decomposer supervision.

tent parsing and are later converted into retrieval and reranking signals over the HSG. The relation labels in `target_relation` are query-side intent labels and are not required to match the HSG edge inventory one-to-one. We report the effective relation-label inventory covered by the decomposer training data: `supported_by`, `emphasizes`, `compares`, `measured_by`, `aligned_to_slide`, `grounded_in_paper`, and `other`. The `other` label serves as a catch-all intent class for questions whose relation is not captured by the core query-side labels.

We train the decomposer with LoRA rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.05. Optimization uses AdamW with a cosine learning-rate schedule, peak learning rate 2×10^{-4}, per-device batch size 4, gradient accumulation steps 8, effective batch size 32, and 3 epochs. The maximum sequence length is set to $1{,}024$, bf16 is enabled, gradient checkpointing is used, and the best checkpoint is selected by validation loss.

Example Decomposer Output

```
{
  "target_semantic_unit": "Result",
  "target_relation": "compares",
  "constraints": {
    "baseline_constraint": "baseline method",
    "metric_constraint": "training efficiency"
  },
  "modality_requirement": "paper+visual",
  "reasoning_pattern": "comparison",
  "answer_type": "explanation",
  "answerable": true
}
```

Table 10: Example ontology-compatible output of the Semantic Decomposer.

F Retrieval and Grounding Configuration

SynapseHSG performs session-scoped multi-channel retrieval over the node index. During inference, the structured constraints produced by the Semantic Decomposer are used as independent query channels against the FAISS index. Retrieved nodes are mapped to candidate triples through an inverted node-to-triple index, and the resulting candidates are merged into a unified recall pool before constraint-aware reranking.

Table 11 reports the retrieval configuration for each channel. The channel weights are unnormalized scoring coefficients used during late-fusion reranking rather than probabilities. The reranker selects the top-10 triples as seed evidence. Hyperedge anchor expansion then adds up to 15 additional triples through offline-retained cross-modal links. Table 12 summarizes the grounding configuration used during inference.

Retrieval Channel	Top-k	Weight
Question	50	0.30
Target semantic unit	20	0.15
Baseline constraint	30	0.25
Metric constraint	30	0.25
Experiment setting	30	0.25
Section constraint	30	0.25

Table 11: Multi-channel retrieval configuration. Weights are scalar scoring coefficients rather than normalized probabilities.

Pipeline Component	Configuration
Offline Indexing	
Dense vector index	FAISS over node embeddings
Candidate expansion	Inverted node-to-triple index
Online Grounding	
Reranker output	Top-10 seed triples
Hyperedge expansion	Up to 15 additional triples
Context Construction	
Typical subgraph size	∼25 nodes / ∼25 triples
Serialized evidence size	∼8,000 chars / ∼2,000 tokens

Table 12: Grounding configuration used during inference.

G Structure-Preserving Linearization Protocol

After reranking and hyperedge expansion, SYNAPSEHSG serializes the grounded sub-HSG into a structured textual context for frozen-LLM generation. The protocol contains two ordered blocks: an evidence-node block and a relation block. Hyperedges are used during traversal to select additional cross-modal evidence, but their memberships are not explicitly printed in the final serialized context.

The evidence-node block lists the selected nodes with their identifiers, modality, node type, source location, and textualized content. The relation block lists induced pairwise relations among the selected nodes. This format preserves local evidence content and graph connectivity while keeping the serialized context bounded.

H Evaluation Protocol

We evaluate model predictions with an LLM-based semantic matching protocol. The main reported metrics, including overall accuracy, answerable accuracy, and refusal accuracy, are computed using GPT-5.4 as the semantic judge. The judge receives the question, reference answer, answerability label, and model prediction, and returns a binary correctness decision according to the rubric in Table 14.

For answerable questions, a prediction is marked correct if it preserves the key facts required by the reference answer, allowing paraphrases and minor surface-form differences. For unanswerable questions, a prediction is marked correct only if it abstains or explicitly states that the available evidence is insufficient. Failed generations and unsupported answers are counted as incorrect.

Block	Serialized Content
Evidence Nodes	Node identifier, modality, node type, source location, and textualized evidence content
Relations	Source node identifier, relation type, and target node identifier

Table 13: Structure of the serialization protocol.

I Baseline Adaptation for Multimodal Evidence

Several iterative and graph-based RAG baselines are originally designed for text-centric corpora and do not natively support heterogeneous conference streams containing papers, slides, and audio. To enable a fair comparison on MM-CONFQA, we adapt their input interface to the multimodal setting while preserving each method's original retrieval, graph construction, or iterative reasoning procedure.

Unified Evidence Interface. We first convert each presentation into a shared pool of textualized evidence units. Paper PDFs are segmented into text chunks. Slides are represented by slide-level textual content, OCR/caption outputs, and localized visual-region descriptions when available. Audio recordings are converted into timestamped transcript segments. These evidence units retain their source metadata, including modality, slide index, page index, and timestamp, but are exposed to text-centric baselines as textual retrieval units.

Adaptation of Text-Centric Iterative RAG. For iterative RAG baselines such as EventRAG, KGGen, and GraphRAG, we use the textualized evidence pool as the input corpus. Their internal retrieval, graph construction, event extraction, or iterative expansion modules are then run without access to oracle evidence paths. In this adaptation, the baselines can retrieve evidence originating

from any modality, but they operate over the textualized representation rather than raw images or audio. This preserves their core algorithmic behavior while allowing them to process the same multimodal source materials used by SYNAPSEHSG.

Adaptation of Multimodal RAG Baselines. For baselines that already support multimodal or document-style retrieval, such as VisRAG, RAG-Anything, and M2RAG, we follow their intended input format whenever possible. When a method requires a specific representation, such as page images, document chunks, OCR text, or captioned visual regions, we provide the corresponding modality-specific evidence derived from the same MM-CONFQA sources.

Controlled Generation and Evaluation. For all retrieval-based and iterative baselines, the final answer generator is fixed to `Qwen3.6-27B` in the controlled setting. Retrieved evidence is concatenated with the original question under the same answer-generation instruction. No baseline receives oracle evidence paths, manually selected contexts, or additional supervision from the test set. All predictions are evaluated using the same GPT-5.4-based semantic judging protocol described in Appendix H.

This adaptation favors the baselines by making slide and audio information retrievable as text, rather than restricting them to paper-only evidence.

J Error Composition Analysis

Figure 7 decomposes total errors into answerable-question errors and unanswerable-refusal errors. This analysis complements the answer-refusal trade-off in the main text by identifying where each method fails. Iterative RAG baselines such as EventRAG and KGGen produce few unanswerable errors, but most of their failures come from answerable questions. In contrast, SYNAPSEHSG substantially reduces answerable-question errors, which are the dominant source of failure in MM-CONFQA. This indicates that the main benefit of SYNAPSEHSG comes from improving grounded answer generation rather than from conservative abstention.

K Backbone Gain Decomposition

Figure 8 decomposes the improvement of SYNAPSEHSG over each native generator into overall accuracy gain, answerable-correct gain, and

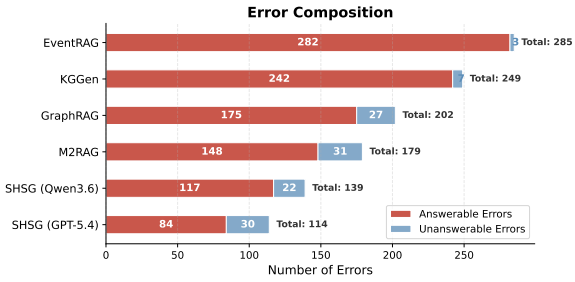

Figure 7: Error composition on the MM-CONFQA test set. Errors are decomposed into answerable-question errors and unanswerable-refusal errors. SYNAPSEHSG reduces the dominant answerable-error component compared with retrieval and iterative baselines.

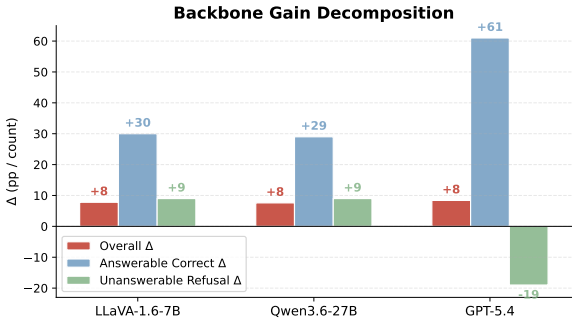

Figure 8: Backbone gain decomposition. Overall gain is measured in percentage points, while answerable-correct and unanswerable-refusal changes are measured by instance count.

unanswerable-refusal change. Across all backbones, SYNAPSEHSG increases the number of correctly answered answerable questions. The effect is especially pronounced with GPT-5.4, where answerable correct predictions increase by 61, while correct refusals decrease by 19. This pattern further supports the conclusion that SYNAPSEHSG primarily improves evidence-grounded answer generation rather than merely increasing abstention.

L Component Sensitivity Profiles

Figure 9 visualizes the per-category performance profiles of the full SYNAPSEHSG model and its ablated variants. Unlike the main ablation plot, which reports aggregate metric changes, this radar visualization shows which reasoning patterns are most affected by each component. Removing anchors yields a clear degradation on multi-hop aggregation and limitation/caveat questions, indicating that anchor-based expansion is important when the answer requires connecting dispersed evidence. Removing the HSG leads to broad degradation, especially on cross-slide and multi-hop categories,

Judge Prompt Template
Given a question, a reference answer, an answerability label, and a model prediction, determine whether the prediction is semantically correct. For answerable questions, mark the prediction correct if it contains the key facts required by the reference answer, allowing paraphrases and minor surface-form differences. For unanswerable questions, mark the prediction correct only if it refuses to answer or states that the evidence is insufficient. Return only `Correct` or `Incorrect`.

Table 14: Prompt template for LLM-based semantic evaluation.

showing the importance of the structured graph index for organizing evidence distributed across slides and modalities.

M Case Study: Cross-Modal Multi-Hop Grounding

We provide a qualitative case study to illustrate how SynapseHSG performs query decomposition, multi-channel retrieval, hyperedge expansion, structure-preserving linearization, and grounded generation. The example is taken from a NeurIPS 2025 presentation, *Alignment of Large Language Models with Constrained Learning*. The question requires comparing performance trade-offs under different constraint thresholds and identifying the metrics used to quantify the gap between learned and optimal policies.

Question.

> How do the experimental results compare the performance trade-offs of the iterative dual-based alignment method under different constraint thresholds, and what metrics are used to quantify the gap between the learned and optimal policies?

Structured Decomposition. The Semantic Decomposer maps the question into ontology-compatible structural constraints:

Field	Value
`target_semantic_unit`	`Claim`
`target_relation`	`compares`
`baseline_constraint`	optimal policy π^*
`metric_constraint`	R-OPT and U-OPT
`experiment_setting`	PKU-SafeRLHF and Anthropic HH-RLHF datasets
`modality_requirement`	`paper+visual`
`reasoning_pattern`	`comparison`
`answer_type`	`explanation`

Table 15: Semantic Decomposer output for the case study.

Retrieval and Expansion. The decomposed constraints activate multiple retrieval channels, including the original question, the baseline constraint, metric constraint, experiment-setting constraint, and target semantic unit. After deduplication, the recall pool contains approximately 160 candidate triples. The reranker selects the top-10 triples, and hyperedge anchor expansion adds cross-modal evidence through retained evidence links. The final grounded subgraph contains evidence from paper text, slides, and audio-derived cues. Although the decomposer predicts `paper+visual` as the required modality combination, hyperedge expansion may introduce additional linked cues, such as audio-derived emphasis, when they are connected to selected anchors.

Evidence Type	Retrieved Evidence
Claim	Definitions of R-OPT and U-OPT as gaps between learned policy π and optimal policy π^*
Method	Iterative dual-based alignment evaluated on PKU-SafeRLHF and Anthropic HH-RLHF
PaperChunk	Lagrangian duality and iterative dualization procedure
VisualRegion	Slide 6 optimization formulation; Slide 7 dual update; Slide 10 threshold-based scatter plots
ProsodyEvent	Speaker emphasis on sub-gradient computation and constrained optimization

Table 16: Representative evidence recovered by topological grounding in the case study.

Linearized Sub-HSG. The grounded sub-HSG is serialized with the structure-preserving linearization protocol. The serialized context contains an evidence-node block followed by a relation block, without explicitly printing hyperedge memberships.

Generated Answer. Conditioned on the question and the serialized sub-HSG, the frozen generator produces the following answer:

> The method evaluates performance by

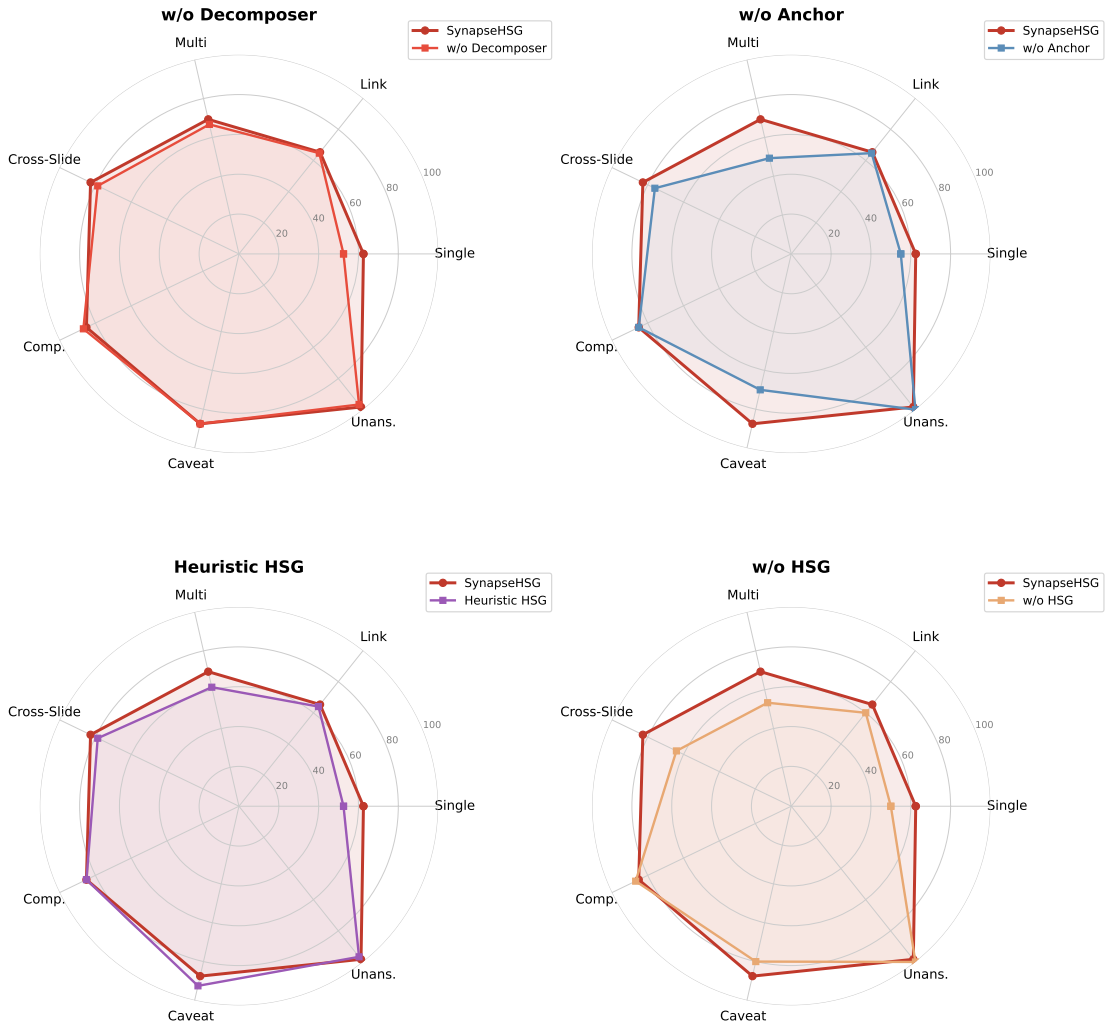

Figure 9: Component sensitivity profiles across reasoning patterns. Each subplot compares the full SynapseHSG model with one ablated variant. Larger enclosed area indicates stronger performance across categories. The radar axes include answerable reasoning categories and unanswerable refusal accuracy.

comparing the learned policy (π) against the optimal policy (π^*) using Reward Optimality (R-OPT) and Utility Optimality (U-OPT) metrics, which measure the expected reward and constraint satisfaction gaps. Experimental results compare these trade-offs across varying constraint thresholds on the PKU-SafeRLHF and Anthropic HH-RLHF datasets, demonstrating how the method balances competing objectives such as harmlessness and humor.

This case illustrates how SynapseHSG connects metric definitions from paper text, optimization and result evidence from slides, and audio emphasis cues into a compact grounded context for multi-hop answer generation.

Linearized Context Fragment

```
=== Evidence Nodes ===
[N1|Method] We demonstrate the effectiveness of our iterative
  dual-based alignment method on PKU-SafeRLHF and Anthropic
HH-RLHF.
[N2|Claim] R-OPT(pi) := |E[br(x,y;pi)] - E[br(x,y;pi*)]|
  U-OPT(pi) := ||E[h(x,y;pi)] - E[h(x,y;pi*)]||_inf
[N3|PaperChunk] We employ Lagrangian duality to develop
  an iterative dual-based alignment method...
[N5|VisualRegion] Slide 6: maximize_theta E[r(x,y)]
  - beta KL subject to E[g(x,y)] >= 0
[N7|VisualRegion] Slide 10: Harmlessness/Humor vs thresholds
[N8|ProsodyEvent] Speaker emphasis on sub-gradient computation
...
=== Relations ===
N1 -[compares]-> N9
N1 -[grounded_in_paper]-> N3
N2 -[grounded_in_paper]-> N4
N1 -[aligned_to_slide]-> N5
N1 -[aligned_to_slide]-> N7
N8 -[emphasizes]-> N1
...
```

Table 17: Linearized sub-HSG fragment used for grounded generation in the case study.