Don't Just Listen, Try Planning: Graph-based Retrieval-Generation Agent for Long-form Audio Meeting Understanding

Quanwei Tang[1], Dong Zhang[1,2*], Shoushan Li[1], and Guodong Zhou[1]

[1]School of Computer Science & Technology, NLP Lab, Soochow University, China
[2]Jiangsu Key Lab of Language Computing, Suzhou.
dzhang@suda.edu.cn

Abstract

While long-form audio meeting understanding (LAMU) is garnering growing attention, task-specific question answering (QA) datasets remain scarce. Existing speech QA paradigms and state-of-the-art Speech LLMs suffer from acoustic information loss and poor long-term context memory. To address these issues, we construct the LongAudioQA dataset and propose the GRGA model, which models heterogeneous audio features into a multi-dimensional graph and leverages agent planning for retrieval and answer generation. GitHub for data and code.

1 Introduction

Long-form audio meeting understanding (LAMU) has attracted significant attention in speech processing. However, previous studies have only focused on transcription recognition for long-form multi-party meetings (Yu et al., 2022a,b; Jain et al., 2024). Consequently, there is a lack of dedicated question answering (QA) datasets for long audio meetings, despite the critical importance of this task. To address this gap, we construct the **LongAudioQA** dataset for LAMU. Distinct from short-form conversation QA, it is designed to capture three core dimensions of long-form audio meeting content: complex semantics, multi-speaker interactions, and quite long timestamps.

Although existing Speech LLMs (KimiTeam et al., 2025; Xu et al., 2025a) have demonstrated strong performance across various tasks, they prioritize textual context by ASR over speech context (You et al., 2022; Lin et al., 2024). Specifically, these models typically map the input speech to the textual context, which inevitably leads to the loss of valuable acoustic information. For example, regarding the query "sudden loud voice at the 19-minute mark" in Figure 1, the inability to access voice volume prevents natural understanding and appropriate response (as the result of Text LLM in the Figure). We denote this phenomenon as **Acoustic Missing**. Additionally, prior QA on speech conversations has been exclusively centered on short-form audio clips or dialogues ($<$ 30s) (Johnson et al., 2024; Zhao et al., 2025; Hu et al., 2026). When applied to extremely long audio meetings, these methods fail to capture long-term dependencies. Let's return to the query in Figure 1, we not only need to locate the content of the 19-minute timestamp, but also to trace the answer back to the description of 1:30 (as our GRGA). Without an explicit design for long-range semantic understanding, even RAG approaches (Tang et al., 2025; Li et al., 2025; Wang et al., 2024) fail to identify the correct rationale (as Text RAG). We designate this phenomenon as **Context Fragmentation**.

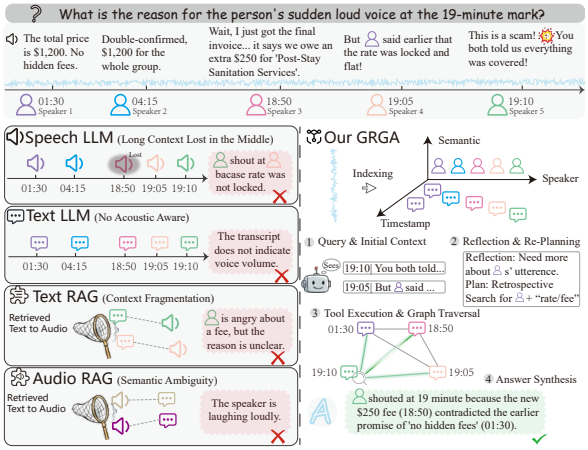

Figure 1: **Existing Speech LLMs and our GRGA for long-form audio QA.** By indexing conversations across **Semantic**, **Speaker**, and **Timestamp** dimensions, our model enables precise reasoning.

To address these challenges, we propose the Graph-based Retrieval-Generation Agent (**GRGA**) model. Specifically, we first model heterogeneous features from the audio, including not only acoustic

*Corresponding Author: Dong Zhang

Findings of the Association for Computational Linguistics: ACL 2026, pages 20715–20742
July 2-7, 2026

information (e.g., voice and tone) but also many speaker attributes (e.g., role: project manager, gender: male), into a unified multi-dimensional graph structure. Then, we leverage agent planning to retrieve question-relevant clues from this multi-dimensional graph and generate the final answer. In summary, we contribute:

• We design and construct the novel dataset **LongAudioQA** for LAMU.

• We propose **GRGA** to handle both acoustic missing and context fragmentation.

• We conduct both automatic and human evaluation on three datasets of our LongAudioQA.

2 Dataset Construction

Existing speech QA benchmarks (Zhifei et al., 2025) predominantly focus on short-context span extraction or simple intent classification. They often treat the dialogue as a flat sequence of text, neglecting the intricate graph-like structure of meeting interactions (e.g., speaker turns, cross-references, and temporal dynamics). Consequently, current models are rarely tested on their ability to perform multi-hop reasoning or temporal grounding over long-form recordings. For instance, answering a question like "How did the speaker's attitude change after the 30-minute discussion?" requires a model to not only localize information but also aggregate evidence across distinct time slices and model the causal dependencies between nodes. Therefore, we propose our **LongAudioQA**.

2.1 Data Selection and Collection

We mainly construct the speech question-answer pairs from the following raw dataset:

AliMeeting. AliMeeting (Yu et al., 2022a,b) is a large-scale Mandarin speech dataset tailored for multi-speaker meeting ASR and SD. A distinguishing feature of this corpus is the simultaneous recording of overlapping far-field audio and individual near-field references. This setup is designed to address the "who said what when" problem, testing the model's ability to handle speaker overlap and diarization in complex settings.

AMI Meeting. The AMI Conference Corpus (Jain et al., 2024) consists of meeting audio recordings captured using far-field microphones, primarily capturing interactions among non-native English speakers. It is characterized by acoustic complexity, presenting significant challenges commonly encountered in real conference settings, such as background noise and reverberation issues.

DailyTalk. In contrast to the aforementioned conference-centric corpora, DailyTalk (Lee et al., 2022) focuses on high-quality open-domain dyadic conversations. This corpus comprises 2,541 dialogues. DailyTalk provides clean acoustic environment with an emphasis on conversational fluency and prosodic features. We concatenated 20 clips to create inputs of intermediate length ($<$10 minutes). While shorter than full meetings, this duration serves to ensure our method performs well on longer audio segments while maintaining effectiveness on shorter ones.

2.2 Question Definition and Data Annotation

To bridge the gap, we constructed a novel dataset designed to expand the boundaries of long-form meeting understanding. Unlike existing research that relies solely on factual retrieval, our dataset introduces a diversified question taxonomy, encompassing factual, inferential, temporal, and acoustic-aware as shown in Table 1). This design enables us to systematically evaluate models' ability to transition from explicit pattern matching to advanced semantic reasoning within complex interaction graphs. Following methodologies for automated data annotation using LLMs (Lian et al., 2025a,b), we employ a hybrid strategy combining large language model-driven automated annotation with human-assisted verification. Specifically, this approach leverages large language models to generate questions from specific local segments of raw conference data based on predefined query types and examples. The models then search conference content for answers and extract detailed evidence. This process is iterative: each batch of generated question-answer pairs undergoes human verification to ensure quality before advancing to the next iteration. While the generation process leverages an "oracle" mechanism to access localized evidence, the core challenge of this study lies in retrieving answers from global, long-context conference data without prior knowledge of relevant segments. This evaluates the model's ability to overcome context fragmentation.

2.3 Dataset Quality Control

Given that automated generation inevitably introduces noise, we enforce a rigorous verification protocol to ensure data reliability. Expert annotators inspect candidate samples to discard hallucinated questions (where answers are absent), rewrite am-

Category	Core Competency	Subtasks & Example Queries
Factual	Explicit Retrieval	• *Entity Retrieval:* "Who mentioned the project code 'Alpha'?" • *Attribute Association:* "What is the budget proposed by the Manager?" • *Keyword Locating:* Matching specific text spans.
Inferential	Multi-hop Reasoning	• *Causal Inference:* Linking a problem raised early with the final decision. • *Coreference Resolution:* Identifying what "that plan" refers to. • *Stance Analysis:* Tracking how a speaker's attitude evolves.
Temporal	Time Awareness	• *Absolute Localization:* "What topic was discussed at the 30-min mark?" • *Relative Sequencing:* "What was discussed *after* 'market research'?" • *Frequency Stats:* Counting term occurrences in a window.
Summarization	Aggregation	• *Topic Summarization:* Synthesizing consensus on "backend architecture". • *Speaker Profiling:* Summarizing a participant's main contributions.
Acoustic-Aware	Multi-modal Alignment	• *Emotion & Intensity:* "Who seemed most **agitated** when discussing the budget?" • *Cross-modal Localization & Causal:* "What is the **reason** for the person's sudden loud voice at the 15-minute mark?"

Table 1: **Taxonomy of Meeting Questions in our LongAudioQA.** Categories are distinguished by background colors: **Factual**, **Inferential**, **Temporal**, **Summarization**, and **Acoustic-Aware**. The **Acoustic-Aware** category uniquely requires grounding textual semantics with paralinguistic acoustic signals.

Dataset	Dur (h)	Avg. Dur (s)	Avg. Turns	# Q&A	# Dial.
AliMeeting	14.91	1,935	815	3,013	28
AMI	18.24	1,930	757	3,243	34
DailyTalk	21.59	610	186	10,200	128

Table 2: Statistical information among DailyTalk, AMI, and AliMeeting datasets. **Dur**: Duration, **Avg**: Average.

biguous references, and validate logic depth (hop count) to ensure complexity balance. Specifically, we require that the generated timestamp evidence yields an Intersection-over-Union (IoU) of > 0.9 with the ground truth. This process results in a high-quality dataset with a human inter-annotator agreement rate of $\kappa = 0.91$ (Cohen's Kappa). Finally, the duration and sample statistics of our LongAudioQA, along with the question distribution, are summarized in Table 2 and Figure 2.

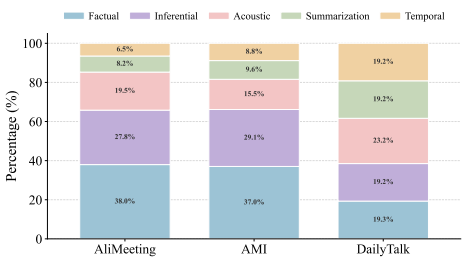

Figure 2: Question Distribution Across Corpora

3 Methodology

The core challenge in long-form audio QA lies in handling composite queries that involve semantic, temporal, and speaker relationships. Existing methods often rely on "one-shot" retrieval, which fails to capture the intricate dependencies in meeting data. To address this, we propose a novel Graph-based Retrieval-Generation Agent (**GRGA**) that mimics the cognitive process of a human expert: it does not merely search, but rather *plans* and *reflects*.

Cognitive Inspiration. To design an effective agent, we draw inspiration from the cognitive strategies human experts employ to navigate complex meetings. We observe that resolving composite queries typically adheres to a **"Search-Reason-Verify"** cognitive loop, which adapts dynamically to the query type. For instance, in *factual* tasks (e.g., verifying budget details), humans execute a targeted keyword scan followed by contextual verification—a mechanism we emulate through retrieval tools. Conversely, *inferential* tasks (e.g., discerning the rationale behind a rejection) necessitates maintaining working memory while traversing causal chains. Similarly, humans naturally construct a *mental timeline* to address *temporal queries*, whereas *summarization* involves synthesizing fragmented details into high-level concepts. Consequently, we architect our agent to replicate these cognitive behaviors by incorporating a **Planner**

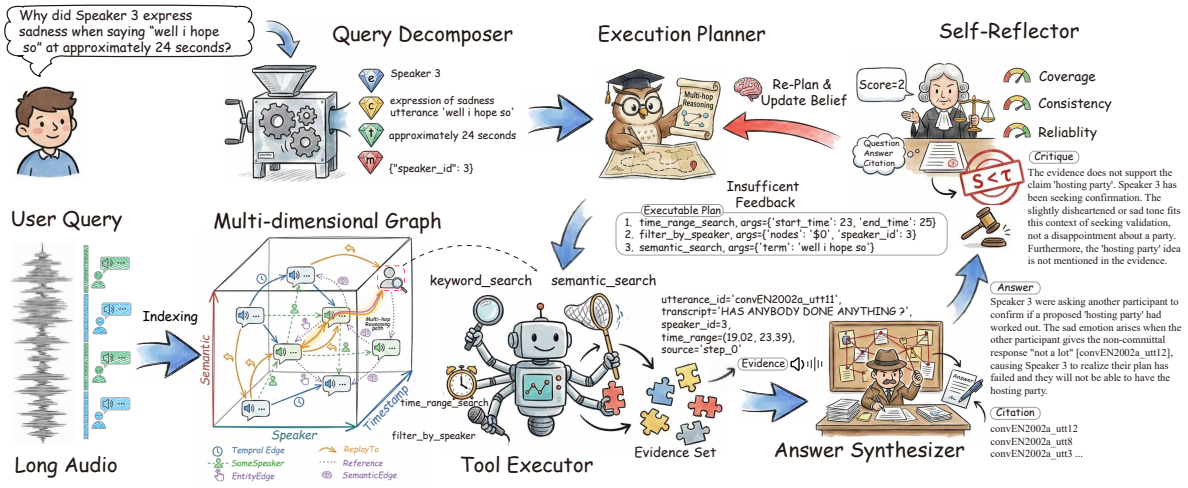

Figure 3: The overall architecture of our graph-based retrieval-generation agent **GRGA**. The framework models long-form audio as a **Multi-dimensional Graph** (bottom-left) to capture semantic, temporal, and speaker dependencies. Our GRGA Planning Process consists of: **Query Decomposition**, **Planning**, **Execution**, **Synthesis**, and **Reflection**.

for logical orchestration, a **Synthesizer** module for holistic information aggregation, and a **Self-Reflector** for answer verification.

3.1 Problem Formulation

We formulate the task as a **Partially Observable Markov Decision Process (POMDP)** (Lauri et al., 2023a), defined by the tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{O}, \gamma \rangle$. The pre-constructed multi-dimensional graph $\mathcal{G}$ serves as the environment with which the agent interacts.

Belief State (b_t) & Policy (π): The belief state b_t, represented by the query Q and interaction history H_t, serves as the sufficient statistic for decision-making. We leverage a pre-trained LLM as the policy $\pi(a_t \mid b_t)$. Implemented via the *Query Decomposer* and *Execution Planner*, the policy utilizes in-context reasoning to determine the next optimal action without parameter updates.

Action ($\mathcal{A}$) & Observation (Ω): The action space $\mathcal{A}$ comprises topological retrieval operations. The *Tool Executor* executes a_t on $\mathcal{G}$, yielding a partial observation $o_t \in \Omega$ (e.g., specific acoustic cues or semantic nodes), which updates b_t.

Reward ($\mathcal{R}$): The *Self-Reflector* acts as a verifier, providing a heuristic reward r_t. It evaluates whether o_t aligns with the reasoning chain, prompting the planner to prune incorrect paths or terminate generation.

Since the model is training-free, our objective is to approximate the optimal trajectory $\tau^* = \{b_0, a_0, \ldots, a_T\}$ that maximizes the reliability of the final answer. The process terminates when the **Answer Synthesizer** determines that sufficient evidence is gathered ($r_t >$ threshold) or the maximum step limit is reached. Figure 3 illustrates our framework.

3.2 Audio-to-Graph Indexing

3.2.1 Acoustic Semantic Alignment

To initialize the graph nodes, we transform the continuous audio stream into discrete, semantically meaningful units with precise temporal boundaries.

VAD and Transcription. We employ FSMN-VAD to segment audio $\mathcal{A}$ into clips, filtering silence. These clips are transcribed by a high-fidelity ASR model to obtain text T. We apply CTC-based Forced Alignment: for each token w_i, we perform constrained Viterbi decoding to align text with acoustic features, yielding exact boundaries $[t_{start}, t_{end}]$.

Speaker Diarization. Identifying "who spoke" is insufficient; understanding "who they are" (e.g., role, stance) is key to reasoning. We propose a multimodal profiling mechanism. We extract speaker embeddings using ERes2NetV2 and perform incremental clustering. If the cosine similarity between a new embedding and existing clusters is below a threshold $\eta = 0.8$, a new speaker ID is created.

Semantic Role Generation. A raw ID (e.g., `Speaker_0`) lacks semantic context. We construct a **Speaker Profile** $\mathcal{P}_k$ for each unique speaker S_k. We aggregate all utterances belonging to S_k and prompt an LLM to distill attributes:

$$\mathcal{P}_k = \text{LLM}(\text{Concat}(\{v_i.\text{text} \mid v_i.\text{spk} = S_k\})) \quad (1)$$

The output is a structured profile, e.g., {Role: Project Manager, Gender: Male, Stance: Conservative}. This profile is stored as a node attribute.

3.2.2 Multi-dimensional Graph Modeling

Beyond the raw meeting transcripts, we also incorporate the acoustic features (e.g., speaker identity, start/end timestamps, and corresponding speech) and explicitly model the dialogue flow as a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$.

Nodes ($\mathcal{V}$): Each utterance is treated as a fundamental node v_i, enriched with attributes including text transcripts, speaker identity, start/end timestamps, and corresponding speech, etc.

Edges ($\mathcal{E}$): To support the diverse reasoning tasks, we construct multiple edge types:

(1) *Temporal Edges (e_{temp})* connecting adjacent utterances ($v_i \rightarrow v_{i+1}$);

(2) *Reply-To Edges (e_{reply}):* Captures conversational turns. We establish a directed edge $v_i \rightarrow v_j$ if v_j occurs within a dynamic window $\Delta t < 5s$ after v_i and speakers are different;

(3) *SameSpeaker Edges (e_{spk})* Connects nodes v_i, v_j where $\text{spk}(v_i) = \text{spk}(v_j)$. This enables the agent to aggregate scattered opinions of a specific person;

(4) *Entity Edges (e_{ent})* derived from entity keyword co-occurrence or discourse relations.

(5) *Semantic Edges (e_{sem})* To solve context forgetting, we employ an LLM to detect coreference chains. If v_j contains pronouns (e.g., "that idea") referring to an entity in v_i, we create a semantic link $v_j \xrightarrow{sem} v_i$.

3.3 GRGA Planning Process

Belief Initialization: Query Decomposition. Given a user query Q, the belief state is initialized by projecting the unstructured query into a structured constraint space $\mathcal{C}$. We define the decomposition function $f_{\text{dec}} : \mathcal{Q} \rightarrow \mathcal{C}$ as:

$$\mathcal{C} = f_{\text{dec}}(Q) = \{c_e, c_c, c_t, c_m\} \tag{2}$$

where c_e, c_c, c_t, c_m denote constraints on **entities**, **concepts**, **time**, and **metadata**, respectively. The initial belief is set as $b_0 = H_0 = \{Q, \mathcal{C}\}$.

Policy Network: Execution Planning. The Execution Planner approximates the policy π_θ parameterized by an In-Context Learning (ICL) (Dong et al., 2024) guided LLM. At step t, it generates a multi-step plan $\mathcal{P}_t$ conditioned on the current belief b_t:

$$\mathcal{P}_t \sim \pi_\theta(\cdot \mid b_t) \tag{3}$$

The plan $\mathcal{P}_t$ is a sequence of atomic operations derived from the Action Space $\mathcal{A}$ (see Table 11):

$$\mathcal{P}_t = [op_1, op_2, \ldots, op_k], \quad op_i \in \mathcal{A} \tag{4}$$

This formulation enables **long-horizon planning**, allowing the agent to chain logical operations (e.g., Search $\rightarrow$ Filter) before interacting with the environment.

Environment Interaction: Tool Execution. The Execution Engine functions as the transition interface. It executes the plan $\mathcal{P}_t$ on the graph $\mathcal{G}$ to obtain an observation o_t:

$$o_t = \text{Exec}(\mathcal{P}_t, \mathcal{G}) = \{s_1, s_2, \ldots, s_n\} \tag{5}$$

where each segment $s_i = (\text{text}_i, \text{time}_i, \text{spk}_i)$. Upon receiving o_t, the agent updates its belief state via a deterministic transition function ψ:

$$b_{t+1} = \psi(b_t, \mathcal{P}_t, o_t) = b_t \oplus \{\mathcal{P}_t, o_t\} \tag{6}$$

Reward Estimation: Synthesis & Reflection. To guide the reasoning trajectory refinement, we employ a two-stage mechanism as reward function.

1) Answer Synthesis. The synthesizer generates a candidate answer $\hat{A}_t$ and citations Cite_t based on the accumulated evidence in b_{t+1}:

$$(\hat{A}_t, \text{Cite}_t) = f_{\text{syn}}(Q, b_{t+1}) \tag{7}$$

2) Reflection as Sparse Reward. The Reflector evaluates the logical entailment between the evidence $E \subset b_{t+1}$ and the answer $\hat{A}_t$, assigning a verification score $s_{\text{ver}} \in [0, 1]$:

$$s_{\text{ver}} = f_{\text{ref}}(Q, \hat{A}_t, E) \tag{8}$$

The reward r_t is defined as a threshold function:

$$r_t = \begin{cases} 1 & \text{if } s_{\text{ver}} \geq \tau \quad \text{(Success, Terminate)} \\ -\beta & \text{if } s_{\text{ver}} < \tau \quad \text{(Failure, Re-plan)} \end{cases} \tag{9}$$

If $r_t < 0$, the negative feedback fb_t (critique) is injected into the belief state: $b_{t+1} \leftarrow b_{t+1} \cup \{fb_t\}$, prompting the policy π to generate a corrective plan $\mathcal{P}_{t+1}$ in the next iteration.

4 Experimentation

To validate the effectiveness of our GRGA, we conduct extensive experiments on our LongAudioQA dataset.

Method	AliMeeting					AMI Meeting					DailyTalk				
	Fact.	Infer.	Temp.	Summ.	Acou.	Fact.	Infer.	Temp.	Summ.	Acou.	Fact.	Infer.	Temp.	Summ.	Acou.
Speech as Context															
Qwen3-Omni	29.31	27.33	15.81	19.06	16.50	20.96	26.38	15.12	16.67	12.15	34.26	38.43	2.43	2.11	9.39
Audio Flamingo 3	35.70	35.23	13.19	18.09	32.89	16.04	27.94	12.62	10.24	17.06	65.13	61.31	3.83	2.44	40.09
MiMo-Audio	54.50	54.50	16.92	39.91	25.15	52.75	64.74	35.34	38.30	17.17	81.57	78.65	5.58	3.46	31.44
Transcription as Context															
Qwen3-Omni	47.66	63.53	38.34	32.65	11.70	53.55	46.66	37.32	43.63	14.13	80.11	77.14	62.71	15.71	18.27
Audio Flamingo 3	32.12	45.16	28.23	21.06	9.69	48.52	40.21	22.12	29.16	12.36	76.43	75.26	58.32	13.42	15.85
MiMo-Audio	48.02	64.62	38.29	29.43	15.72	53.62	44.10	37.14	41.49	17.94	80.67	76.16	62.46	15.35	17.52
Both Speech and Transcription as Context with RAG															
BGE-M3	43.24	29.68	31.56	15.43	18.45	32.56	24.56	23.78	25.36	25.34	46.72	36.31	32.46	34.79	26.21
CLASP	25.37	19.54	18.92	14.76	16.32	24.12	20.31	22.76	11.60	14.79	30.58	33.42	27.34	13.52	11.81
GRGA(ours)	**57.31**	**66.45**	**39.48**	**44.29**	**39.91**	**59.46**	**65.29**	**38.56**	**48.46**	**35.68**	**85.25**	**79.06**	**63.58**	**61.10**	**52.32**

Table 3: **Main results on accuracy (%).** We compare our proposed method against End-to-End Speech LLM and standard RAG baselines across three datasets. **Bold** indicates the best performance. *Fact.*: Factual, *Infer.*: Inferential, *Temp.*: Temporal, *Summ.*: Summarization, *Acou.*: Acoustic reasoning.

4.1 Baselines and Implementation

First, we compare three recent most competitive multi-modal LLMs: Qwen3-Omni (Xu et al., 2025a), **Audio Flamingo 3** (Goel et al., 2025), and MiMo-Audio (Xiaomi, 2025), which are considered as SOTAs for speech or spoken text QA. Therefore, we implement them in two forms: 1) use whole meeting **Speech as Context** and 2) use whole meeting **Transcription as Context**.

Second, we investigate two tailored RAG approaches for audio QA: **BGE-M3** (Chen et al., 2024a) retrieving meeting text through a query, then merging the retrieved text and corresponding audio into the model (TextRAG) and **CLASP** (Abootorabi and Asgari, 2025), which are also considered as the SOTAs for RAG-based audio QA, retrieving meeting audio through a query, then merging the retrieved audio and corresponding text into the model (AudioRAG). They cast both meeting **Speech and Transcription as Context**. Notably, both of them adopt the same LLMs as ours to implement. More details of our GRGA and above baselines can refer to Appendix G.

4.2 Evaluation Metric

To rigorously assess the reasoning capabilities of our model, we report **Semantic Accuracy** across all datasets. Traditional n-gram metrics (e.g., BLEU, Exact Match) often fail to capture the true validity of generated responses, as they penalize correct answers that differ in lexical surface forms from the ground truth. Drawing upon recent methodologies in complex reasoning evaluation (Mishra et al., 2025; Shui et al., 2023), we move beyond rigid string matching and establish a robust, automated evaluation pipeline that prioritizes semantic equivalence and logical soundness.

Specifically, we employ a high-capability LLM (LLM-as-a-Judge) to approximate human-level judgment. Formally, let $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^{N}$ denote the evaluation dataset, where q_i is the query and a_i is the ground-truth answer. Let $\hat{a}_i$ represent the model-generated response. We define a semantic judgment function $\mathcal{J}$, parameterized by an external expert model (e.g., GPT-OSS-120B (OpenAI et al., 2025)):

$$s_i = \mathcal{J}(q_i, a_i, \hat{a}_i) \in \{0, 1\}, \tag{10}$$

where $s_i = 1$ if and only if $\mathcal{J}$ determines that $\hat{a}_i$ entails the same semantic information as a_i, and 0 otherwise. The judge is prompted to disregard stylistic differences and focus solely on factual consistency and the correctness of reasoning. Finally, the Semantic Accuracy is computed as the expectation of correct judgments:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} s_i \times 100\%. \tag{11}$$

This approach ensures a fair comparison by validating whether the model successfully retrieves and reasons over the core knowledge, regardless of its output phrasing.

4.3 Main Results

Table 3 shows the performance comparison of all baselines of our GRGA on our proposed three

Dataset	WER (↓)	DER (↓)	
		Collar = 0 s	Collar = 0.25 s
DailyTalk	0.85	6.42	3.41
AliMeeting-far	20.38	16.80	13.10
AMI-sdm	18.84	17.60	14.90

Table 4: **Error analysis on benchmark datasets.** We report Word Error Rate (WER) and Diarization Error Rate (DER) under strict (0 s) and standard (0.25 s) collar (tolerance) settings. All metrics are reported in percentage (%), and ↓ denotes that lower values are better.

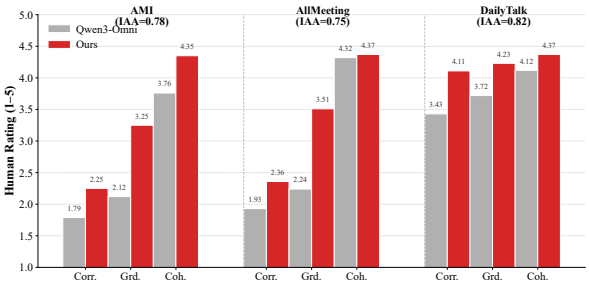

Figure 4: **Human evaluation results.** Our method shows significant improvements, particularly in Groundedness across complex meeting scenarios.

datasets for Long-form speech meeting understanding. From this table, we can see:

Overcoming Context Limitations. End-to-End Speech LLMs (e.g., Qwen3-Omni) degrade sharply on long-form datasets, with AudioFlamingo3 dropping from $\sim$ 65% on DailyTalk to $\sim$ 16% on AMI. This confirms that fixed context windows hinder reasoning in long-form audio. Conversely, our GRGA maintains robust performance, outperforming the strongest baseline (MiMo-Audio) on AMI, validating the scalability of our graph-based retrieval beyond context limits.

Planning vs. Naive Retrieval. Comparisons with RAG baselines highlight the failure of "one-shot" retrieval. Text RAG, while effective for factual questions, struggles significantly with inferential questions (24.6% vs. ours 65.3% on AMI). This proves that vector similarity alone cannot capture multi-hop dependencies. Our Query Planner bridges this gap by decomposing queries into logical chains. Additionally, the poor performance of Audio RAG ($\sim$ 20%) indicates that raw acoustic retrieval is too noisy, justifying our use of a structured graph intermediate.

4.4 Analysis and Discussion

Human Evaluation. As illustrated in Figure 4, our method consistently outperforms Qwen3-Omni

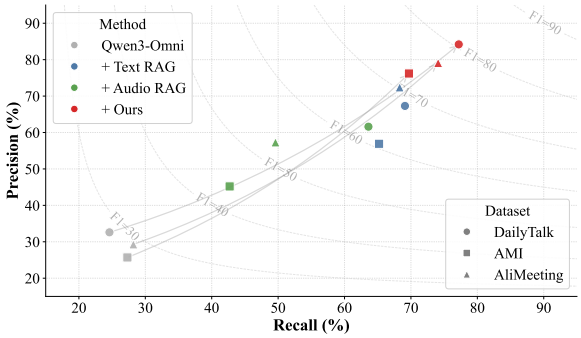

Figure 5: **Citation Precision-Recall Trade-off**. Results are reported in % with a $\pm$2s tolerance.

across all datasets and metrics. Notably, the most substantial gap appears in **Groundedness** (e.g., **+1.27** on AllMeeting). While the baseline generates fluent text (high Coherence), it frequently lacks evidentiary support in multi-party dialogues. Our superior grounding directly translates to higher *Correctness*, confirming that precise temporal anchoring effectively reduces factual errors.

Upstream Performance Analysis. Table 4 highlights the noise disparity across datasets. DailyTalk serves as a clean baseline with minimal errors (WER 0.85%). Conversely, AMI and AliMeeting present high noise rates, with WERs $\sim$ 20% and DERs $>$ 13%. Despite these significant upstream errors in transcription and diarization, our GRGA maintains high QA accuracy (Table 3), demonstrating the robustness of our iterative planning and reflection mechanisms against noisy inputs.

Evidence Recall & Precision Analysis. To assess the model's ability to locate supporting evidence, we report the Precision (P) and Recall (R) of generated citations (timestamps) against ground truth spans (with a $\pm$2s tolerance) in Figure 5 and Table 7. Our GRGA significantly outperforms baselines across all datasets. Specifically, compared to standard Text RAG, our method improves Precision by +19.3% on AMI and +6.7% on AliMeeting. This indicates that our **Query Planner** effectively filters out irrelevant noise, while the **Reflection** mechanism ensures that retrieved segments are strictly relevant to the answer, minimizing hallucinated citations common in naive retrieval approaches.

Ablation Study To validate the necessity of each component in our GRGA, we conduct an ablation study on the AMI dataset as a typical example in Table 5. From this table, we can see that removing any module results in varying degrees of perfor-

Ablation Settings	Overall		AMI				
	Avg.	Δ	Fact.	Infer.	Temp.	Summ.	Acou.
Ours (Full)	**49.49**	-	59.46	65.29	38.56	48.46	35.68
Validation of Query Decomposer							
w/o Query Planner	46.39	-3.10	57.63	59.35	36.79	45.73	32.43
Validation of Query Planner							
w/o Query Planner	44.68	-4.80	56.12	56.35	36.41	43.18	31.37
Validation of Reflection							
w/o Reflection	39.64	-9.84	51.32	49.52	31.10	39.83	26.45
Validation of Action Space							
w/o Tool: Filtering	43.69	-5.79	57.81	62.27	35.36	41.45	28.58
w/o Tool: Semantic Search	16.28	-33.21	28.96	16.38	12.21	13.69	10.15
w/o Tool: Graph Traversal	38.21	-11.27	45.57	38.62	32.31	41.25	33.31
w/o Tool: Audio Access	34.71	-14.78	47.52	49.12	33.13	31.71	12.06

Table 5: **Ablation study on the AMI dataset.** We report the accuracy (%) drop when specific components are removed. Δ: Performance degradation compared to the full framework.

mance degradation for our GRGA. Especially, the results of *w/o Semantic Search* demonstrate that our approach can indeed solve the semantic understanding problem in long-form speech meeting scenario. The removal of *Graph Traversal* indicates the importance of our designed multi-dimensional graph. More analysis can be found in Appendix 6.

Step Analysis can be found in Appendix C. **Case Study** can be found in Appendix D. **Noise Sensitivity Analysis** can be found in Appendix E.

5 Related Work

Speech and Meeting Question Answering. Early research in speech QA primarily focused on extracting answer spans from **short, single-speaker** speech segments, such as Spoken SQuAD (Li et al., 2018) and HeySQuAD (Wu et al., 2023). They normally focus on answering ranking (Hu et al., 2026), instead of generative QA. While recent datasets like AMI (Jain et al., 2024) and AliMeeting (Yu et al., 2022a) provide rich multi-speaker meeting resources, they are predominantly used for ASR and diarization tasks rather than complex reasoning. Existing QA models on these datasets often treat transcripts as flat text sequences, neglecting the intricate temporal and interpersonal dependencies inherent in meetings. In contrast, our work targets **long-form meeting comprehension**, requiring models to navigate graph-structured dialogues involving multiple speakers and temporal dynamics.

Large Speech Language Models. Recent advancements in multi-modal LLMs have enabled direct processing of audio inputs. However, current Speech LLMs face significant limitations in context length. For instance, Kimi-Audio (KimiTeam et al., 2025) is optimized for short clips (< 30s) and suffers from truncation issues with longer inputs. Even state-of-the-art models like AudioFlamingo3 (Goel et al., 2025) are typically constrained to context of approximately 10 minutes. This bottleneck renders them unsuitable for long-form meeting analysis in the absence of external retrieval mechanisms.

Retrieval-Augmented Generation (RAG). RAG has emerged as a standard paradigm for grounding LLMs in external knowledge (Lewis et al., 2020). Traditional "Retrieve-then-Generate" approaches rely on dense vector similarity to retrieve relevant contexts in a single pass. However, these methods struggle with *multi-hop reasoning*, where the evidence is fragmented or requires logical deduction steps not captured by semantic similarity alone (Liu et al., 2025; Tang et al., 2025). Recent works have explored recursive retrieval or chain-of-thought prompting to address these limitations. Our framework advances this paradigm by formalizing retrieval as an **iterative planning process**, specifically tailored to handle the noisy and unstructured nature of ASR transcripts.

LLM Agents and Tool Using. The emergence of LLMs has catalyzed the development of autonomous agents capable of using tools to solve complex tasks, exemplified by frameworks like ReAct (Yao et al., 2023) and Toolformer (Schick et al., 2023). While these agents demonstrate proficiency in open-domain tasks (e.g., web browsing, math) (Zhang et al., 2025), their application to **structured audio understanding** remains underexplored. We bridge this gap by defining a specialized **Action Space** for meeting analysis (e.g., time_range_search, hybrid_search) and incorporating a **Reflection mechanism** modeled as a POMDP (Lauri et al., 2023b), enabling the agent to self-correct in partially observable audio environments.

6 Conclusion

We present **GRGA**, an agentic framework that addresses the acoustic missing and context forgetting issues in long-form speech understanding. By structuring audio into a multimodal heterogeneous graph and formulating QA as a POMDP, we enable an agent to explicitly plan, navigate, and reason over complex interactions. Extensive experiments on our proposed **LongAudioQA** benchmarks demonstrate that our GRGA significantly

outperforms both end-to-end Speech-LLMs and RAG-based SOTAs.

Acknowledgements

This work was supported by Jiangsu Province Frontier Program Project (BF2025036), and Hong Kong RGC grant GRF #15611021. They are also with Jiangsu Key Lab of Language Computing, Suzhou.

Limitations

Our work has three main limitations.

(1) Error Propagation: Since graph construction relies on upstream ASR and diarization, severe acoustic noise or speaker overlap may introduce artifacts into the reasoning graph.

(2) Inference Latency: The iterative agentic workflow (planning and reflection) incurs higher computational cost than single-turn RAG, limiting real-time deployment.

(3) Domain Specificity: Our evaluation focuses on structured meetings; generalizing to unstructured domains like movies or vlogs remains future work.

Ethics Statement

Dataset Sourcing and Compliance. Our dataset is constructed based on three existing public corpora: AliMeeting (Yu et al., 2022a), AMI Meeting Corpus (Jain et al., 2024), and DailyTalk (Lee et al., 2022). We strictly adhere to the original licensing terms of these datasets (e.g., Creative Commons BY-NC-SA 4.0 and Apache License 2.0). We have reviewed the source data to ensure that no Personally Identifiable Information (PII) beyond what was originally consented to by the participants is exposed. For the AMI corpus, we utilize the specific split intended for academic research involving simulated scenarios, minimizing privacy risks associated with real-world private meetings.

Human Annotation and Fair Compensation. We employ highly qualified graduate students as annotators. We ensured that all participants were compensated at a rate significantly above the local minimum wage (approximately $5 per hour), respecting fair labor standards. We also implemented strict protocols to protect annotators from exposure to any potentially harmful or offensive content, although the source datasets are generally free of such material.

Mitigation of LLM Bias and Hallucination. We acknowledge that utilizing Large Language Models (LLMs) for data generation may introduce inherent biases or hallucinations. To mitigate this, our human-in-the-loop pipeline strictly enforces factual consistency checks. We explicitly instructed annotators to discard questions that reinforce stereotypes or rely on hallucinated events not present in the audio. The high inter-annotator agreement ($\kappa = 0.95$) suggests that our verification process effectively filters out low-quality or biased generations.

Potential Societal Impact. The technology proposed in this paper aims to enhance productivity and accessibility by structuring long-form audio. However, we recognize the potential dual-use risk regarding unauthorized surveillance or privacy intrusion in workplace settings. We strongly advocate that the deployment of such meeting analysis tools must be accompanied by transparent consent from all recorded parties and robust data encryption measures. This dataset is released solely for research purposes to advance the field of interpretable audio understanding.

References

Mohammad Mahdi Abootorabi and Ehsaneddin Asgari. 2025. Clasp: Contrastive language-speech pretraining for multilingual multimodal information retrieval. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part IV*, page 10–20, Berlin, Heidelberg. Springer-Verlag.

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

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, Shiliang Zhang, and Junjie Li. 2024b. Eres2netv2: Boosting short-duration speaker verification performance with computational efficiency. *Preprint*, arXiv:2406.02167.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. 2025. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. In *NeurIPS 2025 : Annual Conference on Neural Information Processing Systems*. NeurIPS.

Jiliang Hu, Zuchao Li, Baoyuan Qi, Liu Guoming, and Ping Wang. 2026. End-to-end contrastive language-speech pretraining model for long-form spoken question answering. In *AAAI 2026*.

Aditya Jain, Fagner Cunha, Michael James Bunsen, Juan Sebastián Cañas, Léonard Pasi, Nathan Pinoy, Flemming Helsing, JoAnne Russo, Marc Botham, Michael Sabourin, Jonathan Fréchette, Alexandre Anctil, Yacksecari Lopez, Eduardo Navarro, Filonila Perez Pimentel, Ana Cecilia Zamora, José Alejandro Ramirez Silva, Jonathan Gagnon, Tom August, and 9 others. 2024. Insect identification in the wild: The ami dataset. *Preprint*, arXiv:2406.12452.

Alexander Johnson, Peter Plantinga, Pheobe Sun, Swaroop Gadiyaram, Abenezer Girma, and Ahmad Emami. 2024. Efficient SQA from long audio contexts: A policy-driven approach. In *25th Annual Conference of the International Speech Communication Association, Interspeech 2024, Kos, Greece, September 1-5, 2024*.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, and 21 others. 2025. Kimi-audio technical report. *Preprint*, arXiv:2504.18425.

Mikko Lauri, David Hsu, and Joni Pajarinen. 2023a. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 39(1):21–40.

Mikko Lauri, David Hsu, and Joni Pajarinen. 2023b. Partially observable markov decision processes in robotics: A survey. *IEEE Transactions on Robotics*, 39(1):21–40.

Keon Lee, Kyumin Park, and Daeyoung Kim. 2022. Dailytalk: Spoken dialogue dataset for conversational text-to-speech. *Preprint*, arXiv:2207.01063.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Chia-Hsuan Li, Szu-Lin Wu, Chi-Liang Liu, and Hung yi Lee. 2018. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Preprint*, arXiv:1804.00320.

Qiwei Li, Teng Xiao, Zuchao Li, Ping Wang, Mengjia Shen, and Hai Zhao. 2025. Dialogue-rag: Enhancing retrieval for llms via node-linking utterance rewriting. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24423–24438.

Zheng Lian, Haoyu Chen, Lan Chen, Haiyang Sun, Licai Sun, Yong Ren, Zebang Cheng, Bin Liu, Rui Liu, Xiaojiang Peng, Jiangyan Yi, and Jianhua Tao. 2025a. Affectgpt: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Zheng Lian, Haiyang Sun, Licai Sun, Haoyu Chen, Lan Chen, Hao Gu, Zhuofan Wen, Shun Chen, Siyuan Zhang, Hailiang Yao, Bin Liu, Rui Liu, Shan Liang, Ya Li, Jiangyan Yi, and Jianhua Tao. 2025b. OV-MER: towards open-vocabulary multimodal emotion

recognition. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net.

Chyi-Jiunn Lin, Guan-Ting Lin, Yung-Sung Chuang, Wei-Lun Wu, Shang-Wen Li, Abdelrahman Mohamed, Hung-Yi Lee, and Lin-Shan Lee. 2024. Speechdpr: End-to-end spoken passage retrieval for open-domain spoken question answering. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12476–12480. IEEE.

Hao Liu, Zhengren Wang, Xi Chen, Zhiyu Li, Feiyu Xiong, Qinhan Yu, and Wentao Zhang. 2025. HopRAG: Multi-hop reasoning for logic-aware retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1897–1913, Vienna, Austria. Association for Computational Linguistics.

Venkatesh Mishra, Bimsara Pathiraja, Mihir Parmar, Sat Chidananda, Jayanth Srinivasa, Gaowen Liu, Ali Payani, and Chitta Baral. 2025. Investigating the shortcomings of LLMs in step-by-step legal reasoning. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7795–7826, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b. *Preprint*, arXiv:2508.10925.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

Ruihao Shui, Yixin Cao, Xiang Wang, and Tat-Seng Chua. 2023. A comprehensive evaluation of large language models on legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7337–7348, Singapore. Association for Computational Linguistics.

Quanwei Tang, Sophia Yat Mei Lee, Junshuang Wu, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. 2025. A comprehensive graph framework for question answering with mode-seeking preference alignment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21504–21523, Vienna, Austria. Association for Computational Linguistics.

Mingqiu Wang, Izhak Shafran, Hagen Soltau, Wei Han, Yuan Cao, Dian Yu, and Laurent El Shafey. 2024. Retrieval augmented end-to-end spoken dialog models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 12056–12060. IEEE.

Yijing Wu, SaiKrishna Rallabandi, Ravisutha Srinivasamurthy, Parag Pravin Dakle, Alolika Gon, and Preethi Raghavan. 2023. Heysquad: A spoken question answering dataset. *Preprint*, arXiv:2304.13689.

LLM-Core-Team Xiaomi. 2025. Mimo-audio: Audio language models are few-shot learners.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025a. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Kai-Tuo Xu, Feng-Long Xie, Xu Tang, and Yao Hu. 2025b. Fireredasr: Open-source industrial-grade mandarin speech recognition models from encoder-decoder to llm integration. *Preprint*, arXiv:2501.14350.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Chenyu You, Nuo Chen, Fenglin Liu, Shen Ge, Xian Wu, and Yuexian Zou. 2022. End-to-end spoken conversational question answering: Task, dataset and model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1219–1232, Seattle, United States. Association for Computational Linguistics.

Fan Yu, Shiliang Zhang, Yihui Fu, Lei Xie, Siqi Zheng, Zhihao Du, Weilong Huang, Pengcheng Guo, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022a. M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge. In *Proc. ICASSP*. IEEE.

Fan Yu, Shiliang Zhang, Pengcheng Guo, Yihui Fu, Zhihao Du, Siqi Zheng, Weilong Huang, Lei Xie, Zheng-Hua Tan, DeLiang Wang, Yanmin Qian, Kong Aik Lee, Zhijie Yan, Bin Ma, Xin Xu, and Hui Bu. 2022b. Summary on the ICASSP 2022 multi-channel multi-party meeting transcription grand challenge. In *Proc. ICASSP*. IEEE.

Wenlin Zhang, Xiangyang Li, Kuicai Dong, Yichao Wang, Pengyue Jia, Xiaopeng Li, Yingyi Zhang, Derong Xu, Zhaocheng Du, Huifeng Guo, Ruiming Tang, and Xiangyu Zhao. 2025. Process vs. outcome reward: Which is better for agentic RAG reinforcement learning. *NIPS 2025*.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yu Wang, and Yanfeng Wang. 2025. Librisqa: A novel dataset and framework for spoken question answering with large language models. *IEEE Trans. Artif. Intell.*, 6(11):2884–2895.

Xie Zhifei, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. 2025. Audio-reasoner: Improving reasoning capability in large audio language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23840–23862, Suzhou, China. Association for Computational Linguistics.

A More Analysis in Ablation Study

Ablation Settings	Overall		AMI				
	Avg.	Δ	Fact.	Infer.	Temp.	Summ.	Acou.
Ours (Full)	**49.49**	-	59.46	65.29	38.56	48.46	35.68
Validation of Query Decomposer							
w/o Query Planner	46.39	-3.10	57.63	59.35	36.79	45.73	32.43
Validation of Query Planner							
w/o Query Planner	44.68	-4.80	56.12	56.35	36.41	43.18	31.37
Validation of Reflection							
w/o Reflection	39.64	-9.84	51.32	49.52	31.10	39.83	26.45
Validation of Action Space							
w/o Tool: Filtering	43.69	-5.79	57.81	62.27	35.36	41.45	28.58
w/o Tool: Semantic Search	16.28	-33.21	28.96	16.38	12.21	13.69	10.15
w/o Tool: Graph Traversal	38.21	-11.27	45.57	38.62	32.31	41.25	33.31
w/o Tool: Audio Access	34.71	-14.78	47.52	49.12	33.13	31.71	12.06

Table 6: **Ablation study on the AMI dataset.** We report the accuracy (%) drop when specific components are removed. Δ: Performance degradation compared to the full framework.

To validate the necessity of each component in AudioGraph, we conduct an ablation study on the AMI dataset (Table 6).

Impact of Cognitive Modules. Removing the **Query Planner** leads to a significant drop of 4.80%, particularly in inferential tasks (-8.94%). This confirms that complex queries (e.g., multi-hop reasoning) cannot be solved by single-step retrieval; explicit planning is essential for decomposing intents. Most notably, the removal of the **Reflection** mechanism causes a sharp decline of 9.84%. Without the "verify" loop, the agent is prone to hallucination, accepting the first retrieved chunk even if it is irrelevant. This validates our hypothesis that a POMDP-style feedback loop is critical for robustness.

Impact of Graph Tools. The **Graph Traversal** tool proves indispensable ($\Delta = -11.27\%$). When disabled, the agent degrades to a flat-text searcher, failing to aggregate scattered information via speaker edges (e_{spk}) or temporal edges (e_{temp}). Furthermore, removing **Audio Access** severely impacts acoustic-aware questions (Accuracy 35.68% → 12.06%), demonstrating that text transcripts alone are insufficient for capturing paralinguistic cues like emotion or speaker overlap. Finally, **Semantic Search** serves as the foundational entry point; its removal collapses the system ($\Delta = -33.21\%$), as the agent loses the ability to locate initial evidence nodes.

Method	DailyTalk		AMI		AliMeeting	
	P	R	P	R	P	R
Qwen3-Omni	32.6	24.6	25.7	27.3	29.3	28.2
+ Text RAG	67.3	69.1	56.9	65.2	72.4	68.3
+ Audio RAG	61.6	63.6	45.2	42.7	57.3	49.6
+ Ours	**84.2**	**77.2**	**76.2**	**69.7**	**79.1**	**74.1**

Table 7: **Citation accuracy comparison.** Results are reported in % with a ±2s tolerance. P: Precision, R: Recall.

B Evidence Precision Recall Analysis

Table 7 compares the evidence citation accuracy across three datasets, demonstrating that our method consistently outperforms both the vanilla model and RAG baselines.

C Step Analysis

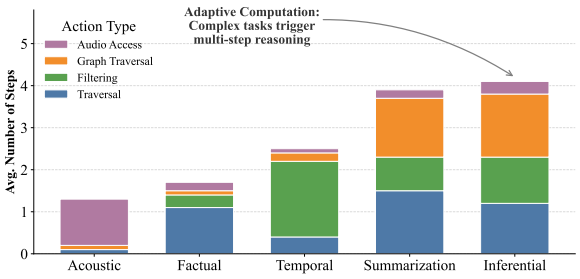

Figure 6: **Average reasoning steps across different question types.** The agent exhibits *adaptive computation*: solving simple Factual queries requires minimal steps, while complex Inferential and Temporal queries trigger deeper reasoning chains. The stacked colors illustrate the distribution of tool usage, showing heavy reliance on **Graph Traversal** for multi-hop reasoning.

Complexity-Aware Reasoning. The results demonstrate a clear correlation between question difficulty and planning depth. For **Factual** queries (e.g., *"What is the budget?"*), the agent adopts a "shortcut" strategy, typically resolving the intent in just 1.8 steps using primarily `Semantic Search`. This indicates high efficiency. In contrast, **Inferential** and **Temporal** queries trigger significantly longer trajectories (avg. >4 steps). This confirms that the agent is actively performing multi-hop reasoning—iteratively traversing e_{temp} or e_{spk} edges to aggregate scattered evidence, rather than relying on a single retrieval pass.

Tool Usage Distribution. The stacked breakdown reveals task-specific tool preferences. **Summarization** tasks show a dominant usage of `Filter` tools (Green) to isolate specific time ranges or speakers. Crucially, **Acoustic-aware** questions exhibit a high frequency of `Audio Access` (Red) calls in the final steps, validating that the agent correctly learns to "listen" to the raw audio only when textual transcripts are insufficient (e.g., verifying emotion).

Dataset	Method	Human Ratings (1-5 Scale)			Avg.	IAA
		Corr.	Grd.	Coh.		
AMI	Qwen3-Omni	1.79	2.12	3.76	2.56	0.78
	Ours	**2.25**	**3.25**	**4.35**	**3.28**	
AllMeeting	Qwen3-Omni	1.93	2.24	4.32	2.83	0.75
	Ours	**2.36**	**3.51**	**4.37**	**3.41**	
DailyTalk	Qwen3-Omni	3.43	3.72	4.12	3.76	0.82
	Ours	**4.11**	**4.23**	**4.37**	**4.24**	
Overall	Average Qwen3-Omni	2.38	2.69	4.07		
	Average Ours	**2.91**	**3.66**	**4.36**	-	-
	Improvement (%)	**+0.52**	**+0.97**	**+0.30**		

Table 8: Human evaluation results across different datasets. Models are assessed on three dimensions: Correctness (Corr.), Groundedness (Grd.), and Coherence (Coh.) using a 1-5 scale. We also report average scores (Avg.) and Inter-Annotator Agreement (IAA).

D Human Evaluation.

To validate real-world performance, we conducted a blind evaluation on 150 queries randomly sampled across five question types from AMI, AliMeeting, and DailyTalk. Three graduate student annotators assessed the responses (IAA=0.78 on avg.). Results in Table 8 shows:

The "Fluency vs. Factualit" Gap. While the baseline (Qwen3-Omni) maintains high *Coherence* (4.07), its low *Groundedness* score (2.69) indicates a tendency to generate fluent but hallucinated content. In contrast, our method achieves a massive **+0.97 improvement in Groundedness**. This confirms that our approach does not merely generate text but accurately anchors answers to precise timestamps and speakers, effectively mitigating hallucinations.

Grounding Drives Correctness. There is a clear positive correlation between groundedness and correctness. By explicitly locating evidence, our method achieves a **+0.52 gain in Correctness**. This demonstrates that superior localization capabilities directly translate to more factually accurate answers, particularly for reasoning-heavy questions.

Robustness in Complex Scenarios. The performance gap is most pronounced on challenging

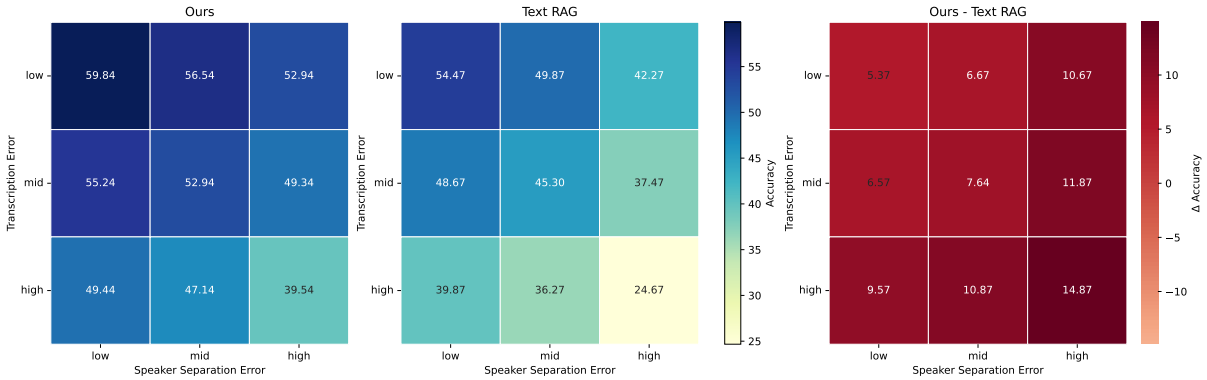

Figure 7: **Noise Sensitivity Analysis.** QA Accuracy comparison under varying upstream error rates. The rightmost panel highlights the widening performance gap (Δ Accuracy), showing that GRGA's advantage grows in noisier environments.

datasets like AMI and AliMeeting (noisy, multi-party interactions) compared to the cleaner DailyTalk. Our method maintains distinct advantages in these "hard" settings (e.g., **+1.13 Groundedness on AMI**), proving its robustness where standard omni-models struggle with speaker attribution and temporal reasoning.

E Noise Sensitivity Analysis

To evaluate the robustness of our proposed GRGA against upstream errors, we conducted a sensitivity analysis by simulating varying levels of Word Error Rate (WER) and Diarization Error Rate (DER). We compare our model with the baseline Text RAG system across a 3×3 grid of noise conditions.

We define three noise levels for both ASR and Diarization components:

Low (Oracle): We use Ground Truth (GT) transcripts and speaker labels to simulate an ideal scenario.

Mid (Standard): We utilize the outputs from our standard pipeline (as described in Section H), representing a realistic deployment scenario (DER $\approx 17.6\%$, WER $\approx 18.8\%$).

High (Simulated Noise): For *High WER*, we utilize a lower-performance lightweight ASR model (openai/whisper-small (Radford et al., 2022)) to generate transcripts with high error rates (WER $\approx 33.6\%$). For *High DER*, we simulated noise by randomly shuffling a subset of speaker labels within a meeting while preserving timestamp boundaries. This represents a worst-case scenario where speaker identity information is highly unreliable (DER $\approx 52.4\%$).

Figure 7 visualizes the impact of noise, revealing distinct behaviors:

Graceful Degradation. Text RAG suffers catastrophic drops as noise increases, falling from 54.47% to 24.67% in the *High/High* setting. This underscores the fragility of standard RAG when explicit text or speaker cues are noisy. Conversely, GRGA demonstrates graceful degradation, retaining 39.54% accuracy even under the harshest conditions.

The "Noise Buffer" Effect. Notably, the performance gap between GRGA and the baseline widens from $+5.37\%$ (Low/Low) to $+14.87\%$ (High/High). This indicates that our graph structure acts as an effective noise buffer. By leveraging topological connectivity and temporal constraints, the agent can compensate for corrupted signals, reducing the system's reliance on perfect transcription and diarization.

F Case Study

Table 9 presents a case study illustrating our method's self-correction capability, where the agent initially generates a low-confidence response but, triggered by the reflection module, re-plans to retrieve specific evidence (e.g., concrete complaints about deadlines), ultimately producing a highly grounded and accurate answer.

G Experimental Setting Details

G.1 Baselines

Qwen3-Omni. Qwen3-Omni (Xu et al., 2025a) is a state-of-the-art (SOTA) end-to-end omni-modal model that processes interleaved inputs (e.g., text and audio) and generates text responses. It supports speech inputs of up to 40 minutes. The model adopts a Mixture-of-Experts (MoE) architecture

Stage	Module / Action	Execution Details & Internal State
User Query: *"What is the team's general attitude to the project deadline and workload?"*		
Phase 1	**Intent Analysis**	**Entities:** ['team', 'project deadline', 'workload'] **Concept:** ['general attitude']
	Plan Execution	**1. Hybrid Search:** query="team project deadline workload" (→ 10 hits) **2. Traverse Relations:** Expand context (→ 64 nodes found).
	Draft Answer	*"The team generally views the deadline as tight... members feeling urgency."* (Generic summary, lacks specific details)
	Reflection (Critic)	**Assessment: ✗Confidence: 2/5** *Reasoning:* "Supported is False. Need reverse confirmation for 'tight' and 'work-intensive' claims." ↪ **Decision:** Trigger RE-PLAN.
Phase 2	**Re-Plan Strategy**	**New Goal:** Validate sentiment strength (e.g., stress, specific complaints). **Tool:** `hybrid_search(query="project deadline focused effort", alpha=0.6)`
	Evidence Retrieval	The new search strategy retrieves highly specific emotional utterances: • `utt180`: "...only got **six days** to do the topic segmentation" • `utt225`: "shall we **change the deadlines**?" • `utt349`: "that's a **hassle** too it's a **lot of work**" • `utt553`: "i wanna do it **superficially**"
	Final Response	*"The team's attitude indicates the project deadline is perceived as tight... One member states they **only have six days** [utt180]. Another asks **'shall we change the deadlines?'** [utt225]. Another calls it **'a hassle'** [utt349]... expresses a preference to handle it **'superficially'** [utt553]."*
	Final Reflection	**Assessment: ✓Confidence: 5/5** **Status:** Supported.

Table 9: **Case Study on Self-Correction.** Initially, the model retrieved broad context but generated a low-confidence (2/5), generic answer. The *Reflection* module detected the lack of specific evidence and triggered a re-planning step focusing on "focused effort." This allowed the model to retrieve concrete complaints (e.g., "six days," "hassle," "superficially"), resulting in a highly grounded (5/5) final response.

with 30B parameters, of which 3B are activated per inference. Notably, it achieves performance comparable to top-tier proprietary models (e.g., Gemini 2.5 Pro (Comanici et al., 2025)) across a wide range of audio understanding benchmarks.

AudioFlamingo3. We select **Audio Flamingo 3** (Goel et al., 2025) as a SOTA audio-language model. Built upon Whisper encoder and 7B LLM, it employs an "on-demand thinking" mechanism to facilitate chain-of-thought reasoning. Crucially, its architecture processes inputs via 30-second windows with a strict context cap of 10 minutes. This baseline serves to quantify the performance degradation of context-constrained models when applied to long meeting scenarios.

MiMo-Audio. MiMo-Audio (Xiaomi, 2025) is a SOTA end-to-end audio language model. It combines a 1.2B audio tokenizer with a 7B LLM. The model supports a 128k context window and downsamples audio tokens to 6.25 Hz, enabling efficient processing of long sequences. Crucially, its instruction-tuning stage incorporates "thinking" mechanisms, allowing chain-of-thought reasoning directly over audio inputs. Comparing against MiMo-Audio helps assess whether our graph-based structural reasoning provides advantages beyond large-scale end-to-end pretraining.

Text RAG Baseline. We implement a standard dense retrieval baseline using **BGE-M3** (Chen et al., 2024a) embeddings. Meeting transcripts are segmented into utterances with prepended metadata (e.g., `[time] SPK: text`). Given a query q, we retrieve the top-k ($k = 10, \lambda = 0.25$) most relevant segments based on cosine similarity, get text result and corresponding audio clips. Crucially, to maintain discourse coherence, retrieved segments are temporally reordered before being fed into the **Speech-LLM**. This baseline represents the conventional approach to handling long meetings without graph-based structural reasoning.

Audio RAG Baseline. We construct a cross-modal retrieval baseline using **CLASP** (Abootorabi and Asgari, 2025), which aligns audio and text in a shared semantic space. The long audio is segmented into clips based on speaker diarization timestamps. Given a text query q, we retrieve the

top-k most relevant audio clips via cosine similarity between the query embedding and CLASP-encoded audio embeddings. These retrieved raw audio clips, along with their speaker metadata, are then fed directly into the **Speech-LLM** (same backbone as ours) to generate the response. This baseline evaluates the efficacy of retrieving raw acoustic features versus our proposed graph-based semantic navigation.

G.2 Evaluation Metric: Semantic Accuracy

To rigorously assess the reasoning capabilities of our model, we report **Semantic Accuracy** across all datasets. Traditional n-gram metrics (e.g., BLEU, Exact Match) often fail to capture the true validity of generated responses, as they penalize correct answers that differ in lexical surface forms from the ground truth. Drawing upon recent methodologies in complex reasoning evaluation (Mishra et al., 2025; Shui et al., 2023), we move beyond rigid string matching and establish a robust, automated evaluation pipeline that prioritizes semantic equivalence and logical soundness.

Specifically, we employ a Strong LLM (LLM-as-a-Judge) to approximate human-level judgment. Formally, let $\mathcal{D} = \{(q_i, a_i)\}_{i=1}^{N}$ denote the evaluation dataset, where q_i is the query and a_i is the ground-truth answer. Let $\hat{a}_i$ represent the model-generated response. We define a semantic indicator function $\mathbb{I}(\cdot)$ parameterized by an external judge $\mathcal{J}$ (e.g., GPT-OSS-120B (OpenAI et al., 2025)):

$$s_i = \mathcal{J}(q_i, a_i, \hat{a}_i) \in \{0, 1\}, \tag{12}$$

where $s_i = 1$ if and only if $\mathcal{J}$ determines that $\hat{a}_i$ entails the same semantic information as a_i, and 0 otherwise. The judge is prompted to ignore stylistic differences and focus strictly on factual consistency and reasoning correctness. Finally, the Semantic Accuracy is computed as the expectation of correct judgments:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^{N} s_i \times 100\%. \tag{13}$$

This approach ensures a fair comparison by validating whether the model successfully retrieves and reasons over the core knowledge, regardless of its output phrasing.

H Implementation Details

All the methods that did not explicitly state the model used Qwen3-Omni (Xu et al., 2025a).

Graph Construction Setup. For the initial audio processing, we utilize the **FireRedASR-AED** (Xu et al., 2025b) model for Automatic Speech Recognition (ASR) to transcribe the raw audio meeting data. To obtain precise word-level timestamps (i.e., t_{start} and t_{end}), we employ the **Montreal Forced Aligner (MFA)**.

Speaker Verification: We use a pre-trained **ERes2NetV2** (Chen et al., 2024b) model to extract speaker embeddings. The cosine similarity threshold for speaker clustering is set to $\eta = 0.8$ based on validation set performance. **Edge Construction:** The temporal window for establishing "Reply-To" edges is set to $\Delta t = 5$ seconds. **Attribute Extraction:** We employ Qwen3-Omni (via vllm API) to extract speaker profiles and resolve coreference chains during the node enrichment phase.

Agent Planning Configuration The core planning mechanism of GRGA depends on specific hyper-parameters governing the exploration-exploitation trade-off: **Backbone Model:** The Policy Network π_θ is parameterized by **Qwen3-Omni**, operated with a temperature of 0 to ensure deterministic reasoning paths. **Reward Function:** As defined in Eq. (9), the self-reflection verification threshold is set to $\tau = 4$ (1-5 Scale). The penalty factor for failed verification is set to $\beta = 0.5$. **Search Constraints:** The maximum depth for the reasoning tree is limited to $K_{max} = 5$ steps. If the agent fails to reach a verified conclusion within these steps, the process terminates and returns the current best candidate.

All experiments were conducted on a server cluster equipped with 8 × Ascend 910B (64GB) GPUs.

H.1 Text RAG Baseline.

To benchmark our method against a standard text-based retrieval pipeline, we implement a dense Retrieval-Augmented Generation (RAG) baseline over meeting transcripts.

Context construction. We segment the meeting transcript into individual utterances and treat each utterance as one retrieval unit. Each unit is formatted with temporal and speaker metadata:

$$u_i = \texttt{[start}_i\texttt{ - end}_i\texttt{] SPEAKER\_ID: text}_i.$$

Dense retrieval. We use BGE-M3 (Chen et al., 2024a) as the embedding backbone. Given a query q, we compute embeddings for the query and each utterance:

$$\mathbf{e}_q = f_{\text{BGE}}(q) \in \mathbb{R}^{1024}, \qquad \mathbf{e}_i = f_{\text{BGE}}(u_i) \in \mathbb{R}^{1024}.$$

We apply L2 normalization to enable cosine-based scoring:

$$\hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \qquad s_i = \cos(\hat{\mathbf{e}}_q, \hat{\mathbf{e}}_i) = \hat{\mathbf{e}}_q^\top \hat{\mathbf{e}}_i.$$

We rank all utterances by s_i and retrieve the top-k segments ($k = 10$) that satisfy a minimum similarity threshold $\lambda = 0.25$:

$$\mathcal{C}_{\text{sim}} = \text{TopK}\Big(\{u_i \mid s_i \geq \lambda\},\, k\Big).$$

Temporal reordering. To preserve discourse coherence, we reorder the retrieved utterances by their start time before forming the final context:

$$\mathcal{C} = \text{SortByTime}(\mathcal{C}_{\text{sim}}).$$

Generation. We concatenate the reordered segments as context $c = \text{Concat}(\mathcal{C})$ and prompt the generator to answer *strictly based on the retrieved context*. We use the same backbone LLM as in our main method for a fair comparison:

$$y = \text{LLM}(q, c).$$

H.2 Audio RAG Baseline.

Given a text query q and a long audio recording A, we build a retrieval-augmented generation (RAG) baseline based on CLASP (Abootorabi and Asgari, 2025), a multilingual audio-text representation model that maps raw speech into a shared 768-dimensional semantic space.

Audio segmenttation.

We segment the long audio A into K shorter clips $\{a_k\}_{k=1}^K$, by speaker diarization.

Embedding computation. We compute the query embedding and the audio clip embeddings using CLASP:

$$\mathbf{e}_q = f_{\text{text}}(q) \in \mathbb{R}^{768}, \tag{14}$$

$$\mathbf{e}_k = f_{\text{audio}}(a_k) \in \mathbb{R}^{768}, \quad k = 1, \ldots, K, \tag{15}$$

where $f_{\text{audio}}(\cdot)$ encodes speech (e.g., via HuBERT and spectrogram encoders with a fusion module), and $f_{\text{text}}(\cdot)$ encodes text into the same representation space.

Retrieval. We rank audio clips by cosine similarity and select the most relevant clip (or top-M clips):

$$s_k = \cos(\mathbf{e}_q, \mathbf{e}_k) = \frac{\mathbf{e}_q^\top \mathbf{e}_k}{\|\mathbf{e}_q\|_2 \, \|\mathbf{e}_k\|_2}, \tag{16}$$

$$k^\star = \arg\max_{k \in \{1,\ldots,K\}} s_k. \tag{17}$$

Optionally, we retrieve $\mathcal{K} = \text{TopM}(\{s_k\}_{k=1}^K)$ for multi-evidence prompting.

Generation. We feed the query and retrieved evidence to a Speech LLM to generate the final response:

$$y = \text{LLM}\big(q, \{a_k\}_{k \in \mathcal{K}}\big), \tag{18}$$

where the evidence is provided, contains retrieval speaker diarization and corresponding audio clips.

I Multi-dimensional Meeting Database Construction Algorithm

Complexity. Let M be the number of VAD clips, and let F_m be the number of acoustic frames in clip m. Let N_w be the total number of word tokens after forced alignment, N_s the total number of diarization segments, and N_u the number of utterances.

ASR inference and **CTC forced alignment** are linear in the input length, costing $\mathcal{O}\Big(\sum_{m=1}^M F_m\Big)$ time. **Speaker diarization** requires embedding extraction over frames plus clustering over segments; in practice this is $\mathcal{O}\Big(\sum_{m=1}^M F_m\Big)$ for feature extraction, and an additional clustering cost by AHC is $\mathcal{O}(N_s^2)$. **Speaker assignment** via overlap matching costs $\mathcal{O}(N_w + N_s)$ using the two-pointer implementation in Alg. 1 (a naive all-pairs matching would be $\mathcal{O}(N_w N_s)$). Building utterances from word streams is $\mathcal{O}(N_w)$, and adding temporal edges in the meeting graph is $\mathcal{O}(N_u)$.

The space complexity is dominated by storing word- and utterance-level annotations, i.e., $\mathcal{O}(N_w + N_u)$, plus the storage overhead of the text and vector indices for utterance retrieval.

Algorithm 1: Multi-dimensional Meeting Database Construction

Input: Long meeting audio X; session id sid; hyper-parameters $\theta = \{\texttt{max\_clip\_len}, \texttt{vad\_pad}, \beta_{\text{sil}}, \alpha_{\text{ovlp}}, K\}$

Output: Structured database $\mathcal{D}$; meeting graph $G = (V, E)$; indices $\mathcal{I}$

$X \leftarrow \textsc{Preprocess}(X)$; // resample/mono/normalize

$\mathcal{C} \leftarrow \text{FSMN\_VAD}(X, \texttt{vad\_pad}, \texttt{max\_clip\_len})$; // clips with absolute spans

Initialize tables Clips, Words, SpkSegs, Utterances;

Initialize graph $G = (V, E)$;

foreach *clip* $c \in \mathcal{C}$ **do** // process each VAD clip

  Clips $\leftarrow$ Clips $\cup\{(sid, c.id, c.t_s, c.t_e, c.wav)\}$;

  $(T, \mathbf{P}) \leftarrow \text{ASR}(c.wav)$; // transcript and CTC posteriors

  $\mathcal{W} \leftarrow \text{CTC\_ForcedAlign}(T, \mathbf{P})$;

  ; // $\mathcal{W} = \{(w_i, a_i, b_i, p_i)\}$ word-level timestamps

  $\mathcal{S} \leftarrow \textsc{Diarize}(c.wav)$; // $\mathcal{S} = \{(spk_j, s_j, e_j, emb_j, conf_j)\}$

  $\mathcal{S} \leftarrow \textsc{ClusterAndSmooth}(\mathcal{S})$; // merge/smooth speaker segments

  SpkSegs $\leftarrow$ SpkSegs $\cup\{(sid, c.id, \mathcal{S})\}$;

  $\widetilde{\mathcal{W}} \leftarrow \textsc{AssignSpeakerToWordsFast}(\mathcal{W}, \mathcal{S}, \alpha_{\text{ovlp}})$;

  Words $\leftarrow$ Words $\cup\{(sid, c.id, \widetilde{\mathcal{W}})\}$;

  $\mathcal{U} \leftarrow \textsc{BuildUtterances}(\widetilde{\mathcal{W}}, \beta_{\text{sil}})$; // group into utterances

  foreach *utterance* $u \in \mathcal{U}$ **do**

    Utterances $\leftarrow$ Utterances $\cup\{(sid, u.id, u.spk, u.t_s, u.t_e, u.text, u.audio\_ref, u.conf)\}$;

    $V \leftarrow V \cup \{u.id\}$; // utterance nodes in G

$E \leftarrow E \cup \textsc{AddTemporalEdges}(\text{Utterances})$; // time adjacency

$E \leftarrow E \cup \textsc{AddSpeakerEdges}(\text{Utterances})$; // same-speaker links (optional)

$E \leftarrow E \cup \textsc{AddEntityOrTopicEdges}(\text{Utterances})$; // optional NER/topic links

$\mathcal{I}_{bm25} \leftarrow \textsc{BuildBM25Index}(\text{Utterances.text})$;

$\mathcal{I}_{vec} \leftarrow \textsc{BuildVectorIndex}(\text{Utterances}, K)$;

$\mathcal{I} \leftarrow \{\mathcal{I}_{bm25}, \mathcal{I}_{vec}\}$;

$\mathcal{D} \leftarrow \{\text{Clips}, \text{Words}, \text{SpkSegs}, \text{Utterances}, \mathcal{I}\}$;

return $\mathcal{D}, G, \mathcal{I}$;

J Retrieval-Generation Agent Algorithm

Algorithm 2: GRGA with Formal Notation

Input: $Q \in \mathcal{Q}, \mathcal{G} = (V, E, X)$
Output: $\hat{A} \in \mathcal{A}$ with $Cite \subseteq V \times T$
// **Decomposition:** $f_{dec} : \mathcal{Q} \to \mathcal{C}$
$\mathcal{C} = f_{dec}(Q) = \{c_e, c_c, c_t, c_m\}$
$b_0 = H_0 = \{Q, \mathcal{C}\}; \quad k \leftarrow 0$
// **POMDP Loop**
while $k < K_{max}$ **do**
  // **Policy:** $\pi_\theta : \mathcal{B} \to \Delta(\mathcal{A}^*)$
  $\mathcal{P}_k \sim \pi_\theta(\cdot \mid b_k)$ where $b_k = b_0 \cup H_k$
  if $\mathcal{P}_k = \perp$ **then**
    return $\perp$
  // **Transition:** $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$
  $o_k = \bigoplus_{op \in \mathcal{P}_k} \text{Exec}(op, \mathcal{G})$ where $o_k \in \mathcal{O}$
  // **Belief Update:** $\psi : \mathcal{B} \times \mathcal{A} \times \mathcal{O} \to \mathcal{B}$
  $b_{k+1} = \psi(b_k, \mathcal{P}_k, o_k) = b_k \cup \{\mathcal{P}_k, o_k\}$
  // **Synthesis:** $f_{syn} : \mathcal{Q} \times \mathcal{B} \to \mathcal{A} \times \mathcal{C}$
  $(\hat{A}_k, Cite_k) = f_{syn}(Q, b_{k+1})$
  // **Reward:** $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$
  $s_{ver} = f_{ref}(Q, \hat{A}_k, o_k) \in [0, 1]$
  $r_k = \begin{cases} 1 & \text{if } s_{ver} \geq \tau \\ -\beta & \text{otherwise} \end{cases}$
  if $r_k > 0$ **then**
    return $(\hat{A}_k, Cite_k)$
  else
    $H_{k+1} = H_k \cup \{(\mathcal{P}_k, o_k, \hat{A}_k, r_k)\}; \quad k \leftarrow k + 1$
return $\perp$

Tool	Time Complexity	Description
`Filter`	$O(\lvert V\rvert + \lvert E\rvert)$	Metadata filter
`Search`	$O(\lvert V\rvert \cdot d_{\text{embed}} + k \log k)$	Vector retrieval + sorting
`GraphTraversal`	$O(\bar{d} \cdot \text{depth})$	BFS ($\bar{d}$: avg degree)
`Temporal`	$O(1)$	Indexed temporal query
`AudioAccess`	$O(1)$	Get audio clip

Table 10: Time Complexity of Graph Operations

J.1 Complexity Analysis

We provide a comprehensive theoretical analysis of the time and space complexity of the proposed GRGA algorithm. We examine each component's computational cost and derive the overall complexity bounds.

J.1.1 Time Complexity Analysis

The total time complexity of GRGA is determined by the iterative POMDP loop and its constituent operations:

$$T_{\text{total}} = T_{\text{dec}} + \sum_{k=0}^{K_{\max}} \left(T_{\text{plan}}^{(k)} + T_{\text{exec}}^{(k)} + T_{\text{syn}}^{(k)} + T_{\text{ref}}^{(k)} \right) \tag{19}$$

where $K_{\max}$ is the maximum number of iterations.

Query Decomposition. The decomposition function $f_{\text{dec}} : \mathcal{Q} \to \mathcal{C}$ involves a single LLM inference:

$$T_{\text{dec}} = O(|Q| \cdot d_{\text{model}}) \tag{20}$$

where $|Q|$ denotes the query length in tokens and d_{model} is the model dimension. This is a one-time operation outside the main loop.

Execution Planning. At iteration k, the policy network π_θ generates a plan conditioned on the belief state b_k:

$$T_{\text{plan}}^{(k)} = O(|b_k| \cdot d_{\text{model}} + L_{\text{plan}} \cdot |\mathcal{A}|) \tag{21}$$

where:

- $|b_k| = O(|Q| + |\mathcal{C}| + k \cdot |o_{\text{avg}}|)$ grows linearly with iteration count
- L_{plan} is the average plan length
- $|\mathcal{A}|$ is the action space size

Tool Execution. The execution engine applies L_{plan} operations on the meeting graph $\mathcal{G} = (V, E)$. The complexity depends on the tool type:

The worst-case execution time is:

$$T_{\text{exec}}^{(k)} = O(L_{\text{plan}} \cdot \max\{|V| \cdot d_{\text{embed}}, |V| \log |V|\}) \tag{22}$$

Answer Synthesis. The synthesizer f_{syn} processes accumulated evidence:

$$T_{\text{syn}}^{(k)} = O\left(\left| \bigcup_{i=0}^{k} o_i \right| \cdot d_{\text{model}} + L_{\text{answer}} \cdot d_{\text{model}} \right) \tag{23}$$

where L_{answer} is the generated answer length. Note that $\left| \bigcup_{i=0}^{k} o_i \right| = O(k \cdot |o_{\text{avg}}|)$ grows with iterations.

Reflection. The reflector f_{ref} evaluates logical entailment:

$$T_{\text{ref}}^{(k)} = O\left((|Q| + |\hat{A}_k| + |o_k|) \cdot d_{\text{model}} \right) \tag{24}$$

Aggregate Complexity

$$T_{\text{total}} = O\Big(K_{\max} \cdot \big[(|Q| + k \cdot |o_{\text{avg}}|) \cdot d_{\text{model}} + L_{\text{plan}} \cdot |V| \cdot d_{\text{embed}} \big] \Big) \tag{25}$$

Under typical conditions where $|V| \gg |Q|$ and $d_{\text{embed}} \approx d_{\text{model}}$, this simplifies to:

$$T_{\text{total}} = O(K_{\max} \cdot |V| \cdot d_{\text{model}}) \tag{26}$$

In practice, $K_{\max} \in [1, 5]$ and early stopping (when $s_{\text{ver}} \geq \tau$) significantly reduces the average iteration count $\bar{K} < K_{\max}$.

J.2 Space Complexity Analysis

The space requirements consist of three main components:

$$S_{\text{total}} = S_{\text{graph}} \tag{27}$$

Graph Storage. The meeting graph requires storage for structure and embeddings:

$$S_{\text{graph}} = O(|V| + |E| + |V| \cdot d_{\text{embed}}) \tag{28}$$

For a typical 30 minutes meeting:

- $|V| \approx 900$ nodes
- $d_{\text{embed}} = 1024$ (BGE-M3)
- Storage: $\sim$2MB (structure) + $\sim$3.5MB (embeddings)

K Tools Details

The table 11 shows the definitions of the atomic tools in our Action Space (A).

Category	Tool Signature	Description & Purpose
Retrieval	`keyword_search(query)`	BM25-based search for precise entity/term lookup.
	`semantic_search(query)`	Dense vector retrieval for abstract semantic concepts.
	`hybrid_search(query, $\alpha = 0.6$)`	Weighted combination of keyword and semantic scores to optimize recall.
Filtering	`filte_time_range(`t_{start}, t_{end}`)`	Retrieves utterances strictly within a specified time window.
	`filte_speaker(nodes, spk_id)`	Filters a set of candidate nodes by speaker identity.
Traversal	`traverse_relations(nodes, depth=k)`	Walks along graph edges (e.g., *Next*, *Reply-To*) to trace dialogue threads for multi-hop reasoning.
Audio	`audio_segment(`t_{start}, t_{end}`)`	Retrieves raw waveform data to ground text in acoustic signals (e.g., for emotion detection).

Table 11: The definitions of the atomic tools in our Action Space ($\mathcal{A}$). The Query Planner invokes these tools to interact with the meeting graph and retrieve evidence.

L QA Examples

Header. The box title `Example # (QA_TYPE, Key=#)` indicates the example index, its question type (e.g., **Factual**), and a unique identifier `Key` used to retrieve the corresponding record in the released JSON files.

Evidence section (lower part). Below the dashed line, we display the **verbatim evidence excerpt(s)** for the referenced utterance IDs. Each excerpt is formatted as: `Evidence [utt_id]: [t_start - t_end] SPEAKER_#: transcript`. Here, `[t_start - t_end]` denotes the absolute timestamps (in seconds) within the meeting, and `SPEAKER_#` is the diarization-based speaker label. This layout makes the supervision explicitly *grounded*: readers can directly verify that the answer is entailed by the cited utterance(s), and systems can be evaluated on both answer correctness and evidence retrieval.

Example 1 (Factual, Key=1)

Question: What specific technical difficulty is Speaker 4 facing regarding XML files?
Answer: Speaker 4 has written code to read and remove the XML, but is struggling to store the data (e.g., into a vector) or display it on the screen using Java.
Evidence (utterance IDs): [513]
Rationale: Speaker 4 explicitly states their current coding roadblock regarding Java vectors and display.

Evidence [513]: [1083.43 - 1100.41] SPEAKER_4: i've been trying to g write something to read the x. m. l. and get rid of it and i can get rid of it but i'm having trouble putting it anywhere else so it will come up on the screen for the moment i haven't managed to put it into a vector or whatever in java to play with it.

Example 2 (Inferential, Key=26)

Question: How did the group decide to resolve the issue of having too many pop-up windows?
Answer: The group discussed using tabs (similar to Mozilla) or toggle buttons within a single window to switch between views like the full transcription and the summary, rather than opening separate windows.
Evidence (utterance IDs): [265, 266, 269, 273, 304, 314]
Rationale: Multiple speakers contribute to the idea of using tabs/toggles to manage content within one window frame to avoid clutter.

Evidence [265]: [657.56 - 663.43] SPEAKER_2: so maybe you can just like choose the s same window for transcription and summary.
Evidence [266] : [661.99 - 664.73] SPEAKER_1: hmm so like have a tab there.
Evidence [269] : [665.32 - 669.42] SPEAKER_2: yeah yeah tabs are nice.
Evidence [273] : [669.91 - 671.05] SPEAKER_4: mozilla style.
Evidence [304] : [697.23 - 703.29] SPEAKER_2: yeah uh change the contents of the same window like from transcription to summary.
Evidence [314] : [719.03 - 725.56] SPEAKER_2: no no it could be like transcription summary like two buttons and you just press on which ever you want.

Example 3 (Summarization, Key=53)

Question: What are the key interface design decisions made during this meeting?

Answer: The team decided to: (1) use buttons instead of right-click menus for speaker characterisation, (2) use tabs or buttons to toggle between transcription and summary in the same window rather than separate windows, and (3) implement right-click menus for topics with options to view all meetings containing that topic.
Evidence (utterance IDs): [76, 84, 88, 158, 167, 266, 269, 314]
Rationale: These utterances capture the major UI/UX consensus points reached through discussion.

Evidence [76]: [207.74 - 220.21] SPEAKER_2: i guess a button button makes a bit more sense 'cause otherwise you don't really know that oh what if i right click now what happens then it's like more if it's visual.
Evidence [84]: [225.08 - 232.04] SPEAKER_4: it's more idiot proof isn't it it's got a button.
Evidence [88]: [229.16 - 231.24] SPEAKER_3: that's true yeah it's more intuitive really isn't it.
Evidence [158]: [399.55 - 402.78] SPEAKER_3: so we could do that in a similar way do it right click as well.
Evidence [167]: [411.28 - 424.13] SPEAKER_3: so we have basically two options of of browsing the meetings is by either um searching and opening individual observations and when then we have the interlinking by right click basically.
Evidence [266]: [661.99 - 664.73] SPEAKER_1: hmm so like have a tab there.
Evidence [269]: [665.32 - 669.42] SPEAKER_2: yeah yeah tabs are nice.
Evidence [314]: [719.03 - 725.56] SPEAKER_2: no no it could be like transcription summary like two buttons and you just press on which ever you want.

Example 4 (Temporal, Key=60)

Question: At what point does the team shift from discussing GUI to discussing the interim prototype?
Answer: Around 810–824 seconds (approximately 13–14 minutes into the meeting).
Evidence (utterance IDs): [361, 363, 365]
Rationale: The topic pivot occurs when Speaker 3 asks what prototype they should aim for.

Evidence [361]: [816.58 - 822.38] SPEAKER_3: and finally the prototype he spoke about what kind of prototype could we produce.
Evidence [363]: [824.27 - 828.76] SPEAKER_3: because i'm i'm just you know i go into the lab and i say right what am i gonna change today.
Evidence [365]: [828.76 - 836.48] SPEAKER_3: you know and it kind of just it just develops i'm not aiming for anything do we wanna aim for something.

Example 5 (Acoustic, Key=*80*)

Question: Why did Speaker 3 express sadness when saying "well i hope so" at approximately 24 seconds?
Answer: Speaker 3 sounded sad when expressing hope that others had done some work, likely in response to Speaker 4's question "has anybody done anything" and Speaker 1's admission "not a lot no", suggesting disappointment about the team's progress.
Evidence (utterance IDs): [13, 11, 12]
Rationale: This combines the sad emotion label from node 13 with the contextual text from surrounding nodes to explain the cause.

Evidence [11]: [19.07 - 23.69] SPEAKER_4: has anybody done anything.
Evidence [12]: [21.47 - 23.29] SPEAKER_1: not a lot no.
Evidence [13]: [24.17 - 25.22] SPEAKER_3: well i hope so.

Current Item: 36

🔮 Prediction Data (Pred)

查看完整数据

Basic Info

Key: 36

Type: inferential

Question

What will the team do with the removed audio processing goal?

Answer

They will mention it as future work in the conclusion of their final report, explaining they didn't have enough time.

Evidence Nodes

Nodes: 137, 140, 149, 152, 153

Reasoning

Speaker 2 suggests 'a good future work thing isn't it stick in the conclusion of the final report' and the team agrees to put it under 'changes since the initial specification'.

✅ Ground Truth (GT)

查看完整数据

Basic Info

Key: 36

Type: inferential

Question

What will the team do with the removed audio processing goal?

Answer

They will mention it as future work in the conclusion of their final report, explaining they didn't have enough time.

Evidence Nodes

Nodes: 137, 140, 149, 152, 153

Reasoning

Speaker 2 suggests 'a good future work thing isn't it stick in the conclusion of the final report' and the team agrees to put it under 'changes since the initial specification'.

📌 Evidence Nodes (Evidence Nodes)

🔮 Pred Evidence Nodes

Node List: [137, 140, 149, 152, 153]

Turn ID	Start	End	Speaker	Text
137	357.38	364.72	SPK_2	it's a good thing to have to say
140	364.72	368.42	SPK_2	should if we had more time th
149	379.35	382.66	SPK_3	but put it under changes since
152	382.66	389.01	SPK_3	and say you know we didn't w
153	385.04	391.36	SPK_2	we don't think there's enough

✅ GT Evidence Nodes

Node List: [137, 140, 149, 152, 153]

Turn ID	Start	End	Speaker	Text
137	357.38	364.72	SPK_2	it's a good thing to have to say
140	364.72	368.42	SPK_2	should if we had more time th
149	379.35	382.66	SPK_3	but put it under changes since
152	382.66	389.01	SPK_3	and say you know we didn't w
153	385.04	391.36	SPK_2	we don't think there's enough

Figure 8: Data Display

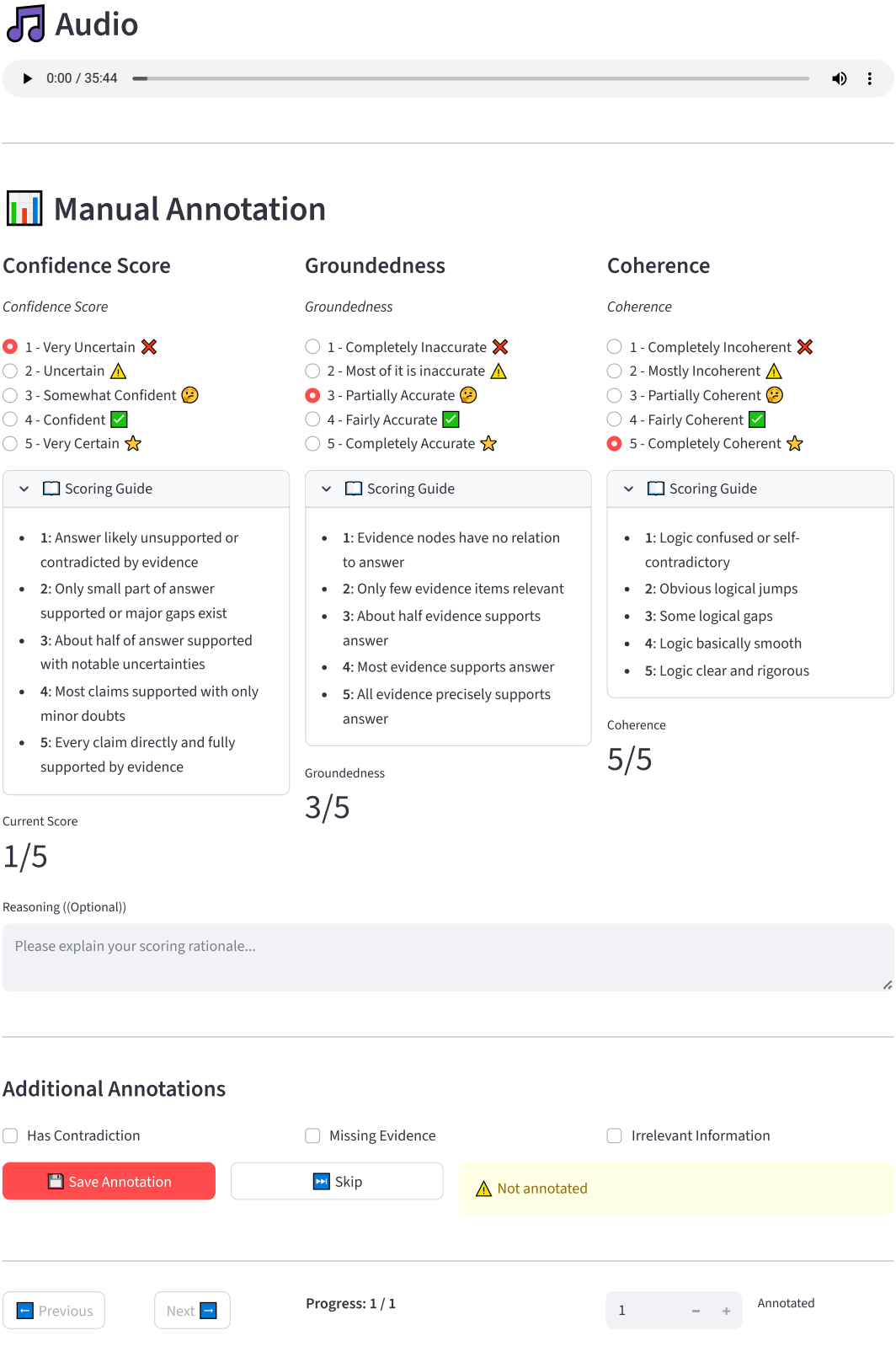

Figure 9: Manual Evaluation

M Manual Evaluation Tool

N Data Construction Prompt

Prompt for Acoustic-Aware QA Generation

 # Role
You are an expert AI assistant specializing in multimodal speech dataset construction. Your task is to generate high-quality **Acoustic-Aware QA pairs** based on the provided audio and speech transcription and metadata.

 # Input Data format
An audio clip.
A list of Nodes. Each Node contains:
‘turn_id‘: [‘timestamp start‘-‘timestamp end‘] Speaker ‘speaker‘: ‘text‘ (Emotion: ‘emotion‘) (Volume: ‘Volume‘)

 # Task Definition: Acoustic-Aware QA
You must generate questions that **cannot** be answered by reading the text alone. The answer must require checking the **Acoustic Features** (specifically the ‘emotion‘ and ‘timestamp‘ fields).

Critical Requirement 1: Focus on High Arousal Emotions
You must prioritize questions regarding strong emotions such as **Anger, Happiness, Excitement, or Sadness**.
How to handle "Neutral" data:
If the input data contains only "Neutral" labels:
- **Do:** Ask **Verification Questions** checking for the presence of strong emotions.
  - Good Example: "Did Speaker 3 sound **angry** when asking about the annual meeting?"
  - Good Answer: "No. According to the audio data, Speaker 3 maintained a **neutral** tone."
- **Don't:** Ask passive descriptive questions like "What was the emotion?".

Critical Requirement 2: NO Confidence Scores
- **Strictly Forbidden:** Do **NOT** mention the numeric confidence scores (e.g., ‘0.99‘, ‘1.0‘) in the Question or the Answer.
- Simply state the emotion label as a fact (e.g., "The speaker was angry").

Allowed Question Patterns
1. **Emotion-Content Attribution (Why is he angry?):**
   "Why did spk3 sound **angry** at the 15th minute?" (Combine Emotion label + Text analysis).
2. **Specific Time Identification:**
   "Who sounded the most **excited** around 10 seconds?"
3. **Emotion Verification (For Neutral Data):**
   "When Speaker 1 said [Text], did they express **happiness**?"
   "Did the speaker sound **furious** or **annoyed** during the discussion?"

NOT Allowed Question Patterns
- **Wrong Questions:**
  Q: Why did Speaker 3 sound sad when saying 'i think the search works as well' around 45 seconds?
  A: Speaker 3 did not sound sad at that moment.

 # Output Format (Strict JSON)
Output a valid JSON List. Ensure ‘evidence_nodes‘ is a list of integers. Using ' instead of " for citation

```json
[
  {{
    "question": "The specific 'question' string.",
    "answer": "The brief 'answer' derived from the text.",
    "evidence_nodes": [10, 12], // Must be a list of turn_ids
    "reasoning": "Brief explanation of the logic."
  }}
]
```

 # Input:

{input_data}

 # Output:

Prompt for Graph-based QA Generation

 # Role
You are an expert data annotator specialized in **Graph-based Audio Understanding** and **Conversational AI**. Your goal is to construct a massive, high-quality Graph RAG dataset from raw ASR transcripts.

 # Input Data Structure
A list of Nodes. Attributes:
‘turn_id‘: [‘timestamp start‘-‘timestamp end‘] Speaker ‘speaker‘: ‘text‘

 # Mission
Your mission is to **exhaustively** mine the provided transcript for every possible piece of information and convert it into a Question-Answer (QA) pair. **Do not stop at just one question per category.** Aim to generate as many valid, distinct QA pairs as the data supports (e.g., 10-20 pairs for a medium-length segment).

 # Critical Guidelines (The "Don't"s)
1. **Filter Noise:** Do NOT generate Factual questions about "YEAH", "OKAY", "MM-HMM" (backchannels).
   - Bad: "Who said 'Yeah'?"
   - Good (Inferential): Use "Yeah" nodes only as evidence for "Agreement" or "Consensus".
2. **Avoid Hallucination:** Every answer must be strictly supported by the ‘evidence_nodes‘.
3. **No Generic Questions:** Avoid vague questions like "What happened?". Be specific: "What software bug did Speaker A mention?"

 # Question Generation Strategy (How to generate MANY questions)
To maximize the number of QAs, apply these specific strategies:

Strategy A: Entity-Centric Mining (Factual)
Scan for **every** entity in the text. Generate a question for each.
- **Entities:** Project names, specific numbers, technical terms, people's names, locations, file paths.
- Trigger: "I see a phone number '04555'." -> QA: "What is the phone number mentioned?"

Strategy B: Logic & Linkage (Inferential)
Look for **connections** between distant or adjacent nodes.
- **Cause & Effect:** Node X proposes something -> Node Y rejects it. -> QA: "Why was the proposal in Node X rejected?"
- **Clarification:** Node A asks a question -> Node B answers it. -> QA: "How did Speaker B respond to A's inquiry about [Topic]?"
- **Sentiment:** Look for emotional words or emphatic language. -> QA: "Which speaker expressed frustration about the file system?"

Strategy C: Time Anchors (Temporal)
- **Absolute:** "What topic was introduced exactly at the 10-minute mark?"
- **Relative:** "What was discussed immediately before the discussion about the budget?"
- **Duration:** "How long did the debate about 'UI Design' last?" (Calculate from start/end timestamps of the cluster).

Strategy D: Scene Understanding (Summarization)
- **Topic Segmentation:** Identify where the topic shifts. -> QA: "Summarize the main points discussed regarding [Topic Name]."
- **Speaker Role:** "Based on the segment, what seems to be the role of Speaker 0?" (e.g., Manager, Technical Lead).

 # Output Format (Strict JSON)
Output a valid JSON List. Ensure ‘evidence_nodes‘ is a list of integers.

```json
[
  {{
    "type": "factual | inferential | temporal | summarization",
    "question": "The specific question string.",
    "answer": "The precise answer derived from the text.",
    "evidence_nodes": [10, 12], // Must be a list of turn_ids
    "reasoning": "Brief explanation of the logic."
  }}
]
```

Input:

{input_data}

Output: