AIMPACT: AN EVIDENCE-GROUNDED MULTI-AGENT FRAMEWORK FOR AUTOMATED PRESENTATION ASSESSMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated presentation assessment is essential for scalable presentation training and feedback in higher education and workplace settings. Despite recent advances in Multimodal Large Language Models (MLLMs), presentation assessment remains constrained by the direct mapping of heterogeneous multimodal evidence to final assessments, which couples evidence reliability with rubric-based judgment. To address this problem, we introduce **AIMPACT**, an evidence-grounded multi-agent framework for automated presentation assessment. Specifically, we first propose an **E**vidence-to-**J**udgment **M**ulti-**A**gent Framework (**EJMA**) to organize presentation assessment into a two-level hierarchy. At the first level, multimodal observations are converted into a traceable evidence memory and coordinate seven dimension-specific specialist agents through a reliability-guided collaboration graph, forming an evidence-grounded council prior. On this basis, the second level introduces **R**ubric-**C**onditioned **M**ulti-**E**xpert **P**reference **O**ptimization (**RC-MEPO**), which first learns expert corrections to the council prior and then performs on-policy preference optimization against the confidence-weighted expert score distribution, where controlled rubric perturbations further enforce assessment consistency with the provided scoring rubric. In addition, we construct a **M**ulti-**E**xpert **P**resentation **A**ssessment Dataset (**MEPA**) that comprises 35 graded classroom presentations with synchronized audio, video, and slides. Five expert instructors independently assessed each presentation across eight rubric dimensions, resulting in 1,400 individual expert judgments. Extensive objective and subjective evaluations show that our method achieves the strongest presentation-level expert alignment among comparable-scale MLLMs and remains competitive with substantially larger proprietary models. Zero-shot evaluations on out-of-domain datasets and additional analyses further demonstrate its ability to generalize to unseen presentation data and different scoring standards. The paper website is available at https://anonymous.4open.science/w/AIMPACT_21373/.

1 INTRODUCTION

Presentation assessment plays a central role in communication training across educational and professional scenarios, where reliable evaluation requires joint consideration of verbal content, vocal delivery, nonverbal behavior, and visual materials. Developing reliable automated methods for such multi-dimensional assessment is therefore important for supporting scalable presentation training and feedback.

As summarized in Figure 1, existing automated presentation assessment approaches evolved through two main paradigms. Earlier feature-based systems extract hand-crafted prosodic, lexical, facial, gaze, gesture, and body-motion cues and map them to presentation scores using task-specific predictors (Wörtwein et al., 2015; Ramanarayanan et al., 2015; Chen et al., 2016). Although this paradigm established the foundation for automated presentation assessment, its dependence on fixed feature spaces and separately engineered scoring models limits its ability to accommodate richer multimodal evidence and varying assessment standards. The recent paradigm has increasingly shifted toward direct MLLM evaluators, where speech, visual, and textual inputs are processed within a unified model, and recent LLM-based judges, rubric-conditioned scoring, and multi-agent evaluation further extend the flexibility of the scoring process (Lepp et al., 2020; Tun et al., 2023; Zheng

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

et al., 2023; Ochoa & Zhao, 2024; Kim et al., 2024a; Chan et al., 2024). Despite this progression, a central methodological limitation remains. Existing multimodal assessment methods still commonly map heterogeneous evidence directly to final scores without separating evidence reliability from rubric-based judgment, which couples evidence acquisition with assessment and makes scoring errors difficult to trace. This challenge is further compounded by limited multi-expert supervision, where existing presentation datasets rarely provide independent expert annotations across multiple assessment dimensions (Park et al., 2014; Chen & Lee, 2017; Lepp et al., 2020; Zhang et al., 2021), limiting expert-aligned learning and evaluation against human expert panel consensus. As a result, the development of more reliable and expert-aligned frameworks for automated presentation assessment remains underexplored.

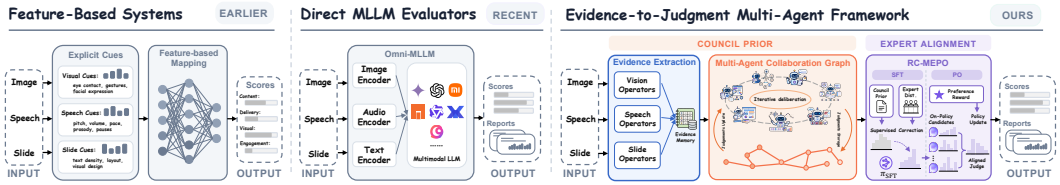

Figure 1: Comparison of three paradigms for automated presentation assessment. Feature-based systems rely on engineered multimodal cues and task-specific predictors, whereas direct MLLM evaluators map multimodal inputs directly to final judgments. Our proposed evidence-to-judgment paradigm instead separates multimodal evidence construction from rubric-based judgment.

To address this problem, we introduce **AIMPACT**, an evidence-grounded multi-agent framework for automated presentation assessment. Specifically, we first propose an **E**vidence-to-**J**udgment **M**ulti-**A**gent Framework (**EJMA**) to organize presentation assessment into a two-level hierarchy. At the first level, multimodal observations are organized into a traceable evidence memory that preserves the observed value, temporal or slide location, confidence, and provenance of each evidence item. A reliability-guided collaboration graph then coordinates seven dimension-specific specialist agents according to evidence reliability, retaining unresolved disagreement as uncertainty and forming an evidence-grounded council prior. On this basis, the second level introduces **R**ubric-**C**onditioned **M**ulti-**E**xpert **P**reference **O**ptimization (**RC-MEPO**), which first learns expert corrections to the council prior and then performs on-policy preference optimization against the confidence-weighted expert score distribution, where controlled rubric perturbations further enforce consistency with the supplied scoring rubric. In addition, we construct a **M**ulti-**E**xpert **P**resentation **A**ssessment Dataset (**MEPA**) that comprises 35 graded classroom presentations totaling 2.81 hours of synchronized audio, video, and slides. Each presentation is independently reviewed by five expert instructors across eight rubric dimensions, yielding 1,400 dimension-level judgments with a quality score, confidence rating, and written feedback. Extensive evaluations across in-domain assessment, out-of-domain assessment, and complementary human and diagnostic analyses show that our method achieves the strongest presentation-level expert alignment among comparable-scale omni-MLLMs and remains competitive with substantially larger proprietary models, while maintaining robust generalization to unseen presentation data and scoring standards.

Our contributions are the following.

- We propose an **E**vidence-to-**J**udgment **M**ulti-**A**gent Framework (**EJMA**) for two-level presentation assessment, where the first level combines traceable multimodal evidence with a reliability-guided collaboration graph to form an evidence-grounded council prior, while the second level introduces on-policy preference optimization to align this prior with the confidence-weighted expert score distribution.
- We construct a **M**ulti-**E**xpert **P**resentation **A**ssessment Dataset (**MEPA**) comprising 35 presentations annotated by 5 expert instructors across 8 rubric dimensions, where each expert provides dimension-level scores, confidence ratings, and written feedback for expert-aligned presentation assessment.
- Extensive objective and subjective evaluations show that our method achieves the strongest presentation-level alignment performance, while evaluation on out-of-domain datasets and additional analyses further demonstrate its robust generalization capability to unseen presentation data and scoring standards.

2 Related Work

Presentation Assessment Methods. Automated presentation assessment has traditionally followed a feature-based paradigm, where prosodic, lexical, gaze, pose, facial, and slide-related cues are explicitly extracted and mapped to scores through task-specific predictors (Wörtwein et al., 2015; Ramanarayanan et al., 2015; Ochoa et al., 2018; Ochoa & Zhao, 2024). While these systems established interpretable multimodal assessment pipelines, their dependence on predefined features and task-specific scoring models limits their flexibility across different assessment settings. More recently, general-purpose foundation-model evaluators have explored how large language models can reason over heterogeneous inputs and scoring rubrics outside the presentation-assessment domain. Representative approaches include MLLM-as-a-Judge (Chen et al., 2024) for multimodal evaluation, Prometheus-Vision (Lee et al., 2024) and Prometheus 2 (Kim et al., 2024b) for fine-grained evaluation under user-defined criteria, and ChatEval (Chan et al., 2024) for multi-agent deliberation among evaluators. More recent studies further model human evaluation uncertainty through distributions over possible judgments rather than collapsing supervision into a single rating (Chen et al., 2025; Wang et al., 2025). However, these advances mainly focus on how final judgments are generated or aligned, while foundation-model-based presentation assessment still commonly couples heterogeneous multimodal evidence with rubric-based scoring decisions, leaving evidence reliability insufficiently separated from scoring decisions and making scoring errors difficult to trace.

Presentation Assessment Datasets. Existing presentation assessment datasets have progressed from holistic performance ratings toward multimodal and dimension-specific annotations. Early datasets combine audio-visual recordings, slides, and human judgments for rubric-based assessment of public-speaking performance (Chen et al., 2014; Wörtwein et al., 2015; Chen et al., 2016), while subsequent corpora extend supervision to attributes such as persuasiveness, audience response, and broader public-speaking skills (Park et al., 2014; Chen & Lee, 2017; Biancardi et al., 2025). Recently, more datasets provide finer-grained supervision for individual communication dimensions or classroom presentation assessment (Lepp et al., 2020; Zhang et al., 2021; Xiong et al., 2025; Becerra et al., 2026). However, existing presentation datasets still rarely provide independent assessments from multiple experts and jointly retain per-rater scores, confidence, and written feedback across a broad presentation rubric, limiting the study of both panel consensus and systematic variation among human expert judgments.

3 Multi-Expert Presentation Assessment Dataset

Data Collection. We construct a **M**ulti-**E**xpert **P**resentation **A**ssessment Dataset (**MEPA**) that comprises 35 graded classroom presentations collected in a university course, where students deliver prepared presentations with slides under a common assessment standard. Each presentation lasts 3.6-6.4 minutes, yielding 2.81 hours of complete classroom recordings. All presentations are captured under the same setting with two 4K video views, including a full view of the presenter and projected slides and a closer view of the presenter, together with a synchronized wireless-microphone audio track. Original slides are retained for all presentations, comprising 380 slides in total and ranging from 6 to 24 slides per presentation. **Expert Annotation.** Before annotation, the recordings are de-identified and the multimodal records are checked for completeness. Five experienced instructors from a university English communication center independently review every presentation with replay permitted, and assess 8 rubric dimensions covering language, pronunciation, volume and pace, fluency, nonverbal delivery, slide design, speech-slide synchronization, and slide reference. For every dimension, each instructor provides a five-level quality score, a confidence rating, and detailed written feedback, resulting in 175 complete expert reviews and 1,400 dimension-level judgments. **Quality Control.** To improve annotation consistency, all experts follow a shared rubric and standardized scoring protocol, with calibration conducted before formal annotation. More details of the annotation protocol and dataset statistics are provided in Appendix A.

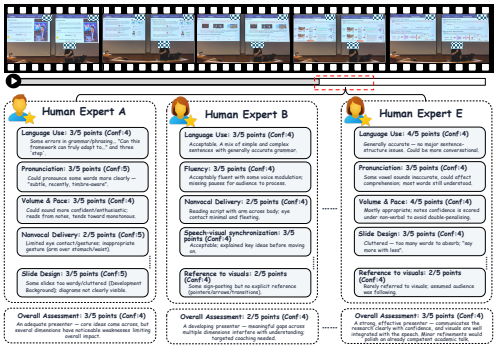

Figure 2: Representative multi-expert annotation example with dimension-level scores, confidence ratings, and written feedback.

4 EVIDENCE-TO-JUDGMENT MULTI-AGENT FRAMEWORK

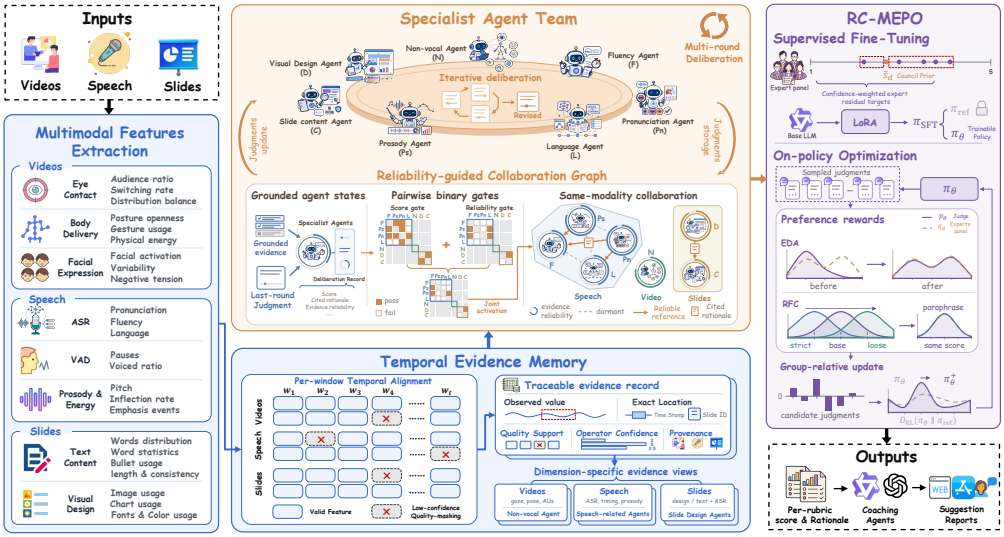

Figure 3: Overview of the proposed two-level evidence-to-judgment framework. The first level constructs an evidence-grounded council prior by organizing multimodal observations into traceable evidence and performing reliability-guided multi-round deliberation among specialist agents. The second level applies RC-MEPO to align the council prior with the confidence-weighted expert score distribution under the supplied scoring rubric. The blue, orange, and purple regions denote Multimodal Evidence Extraction, the Multi-Agent Collaboration Graph, and RC-MEPO, respectively.

4.1 ASSESSMENT TAXONOMY

We propose the **E**vidence-to-**J**udgment **M**ulti-**A**gent Framework (**EJMA**) to organize multimodal presentation assessment into a two-level hierarchy. As shown in Figure 3, the first level constructs an evidence-grounded council prior through two consecutive components. **Multimodal Evidence Extraction** first extracts modality-specific signals from synchronized multimodal inputs, and then organizes the resulting observations into a traceable temporal evidence memory with dimension-specific views. **Multi-Agent Collaboration Graph** then grounds seven specialist agents on these evidence views and coordinates their judgments according to evidence reliability, resolving cross-agent disagreement while retaining unresolved cases as uncertainty to form the council prior. On this basis, the second level applies **RC-MEPO** to learn expert corrections and align the resulting judgments with the confidence-weighted expert score distribution under the supplied scoring rubric. The following sections describe the two components of council prior construction and the subsequent multi-expert preference optimization, respectively.

4.2 COUNCIL PRIOR CONSTRUCTION

4.2.1 MULTIMODAL EVIDENCE EXTRACTION

Directly mapping heterogeneous multimodal signals to rubric-based scores entangles measurement noise with scoring decisions, making it difficult to distinguish weak evidence from incorrect assessment. To address this problem, we separate multimodal evidence construction from subsequent rubric-based judgment. Specifically, synchronized audio, video, and slides are first processed by frozen pretrained models for word-aligned speech recognition (Radford et al., 2023), gaze estimation (Abdelrahman et al., 2023), pose tracking (Jiang et al., 2023), and facial action detection (Chang et al., 2024), together with deterministic multimodal operators that measure speech timing, pronunciation confidence, prosodic variation, slide design, and speech-slide alignment. These modality-specific observations are temporally aligned within presentation windows and stored in a temporal evidence memory. Within each window, low-confidence observations are masked according to their estimated confidence, after which the remaining valid observations are organized into traceable evidence records that retain the observed value, temporal or slide location, operator confidence, and source provenance. These records are then grouped according to the assessment dimensions to form dimension-specific evidence views for subsequent multi-agent scoring.

4.2.2 MULTI-AGENT COLLABORATION GRAPH

Building on the dimension-specific evidence, seven specialist agents assess fluency, prosody, pronunciation, language, non-vocal delivery, visual design, and slide content using only the evidence relevant to their respective assessments. The slide content specialist operates under separate speech-slide synchronization and slide-reference rubrics, allowing the seven agents to cover all eight scoring dimensions. Each specialist first produces an initial score s_d together with an evidence-grounded rationale linked to its evidence set E_d, defining the first-round dimension-specific agent states. Although this restricted evidence access preserves dimension-specific evidence grounding, independent scoring can still produce substantial cross-dimensional score differences under uneven evidence reliability. Therefore, we introduce a reliability-guided collaboration graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ comprises the dimension-specific specialist states and $\mathcal{E}$ represents the active collaboration edges among specialists associated with the same modality. For each dimension, we estimate the reliability ρ_d of its supporting evidence as:

$$\rho_d = \frac{1}{|E_d|} \sum_{e \in E_d} c_e \left(1 - \exp\left(-\frac{n_e}{\tau_e}\right)\right), \tag{1}$$

where n_e denotes the number of valid observations supporting evidence e, c_e is the confidence reported by its multimodal operator, and τ_e controls how rapidly additional observations increase reliability. Based on this, the initial scores and evidence reliabilities determine whether a same-modality agent group enters deliberation. For each modality group m, Δ_m^s, Δ_m^ρ, and ρ_m^* denote its score divergence, reliability imbalance, and strongest evidence reliability, respectively. The collaboration gate is defined as:

$$g_m = \mathbf{1}\big[\exists\, i, j \in m : \underbrace{|s_i - s_j| \geq \tau_s}_{\text{Are scores divergent?}}, \underbrace{|\rho_i - \rho_j| \geq \tau_\rho}_{\text{Is reliability uneven?}}, \underbrace{\max(\rho_i, \rho_j) \geq \tau_g}_{\text{Is reliable evidence present?}}\big], \tag{2}$$

where τ_s, τ_ρ, and τ_g denote the corresponding activation thresholds. Collaboration is activated only when substantial score divergence coincides with uneven evidence reliability and sufficiently reliable supporting evidence.

For each activated group, the specialist agents deliberate over their current scores, evidence reliabilities, and evidence-grounded rationales. Agent states supported by more reliable evidence provide stronger references during deliberation, while each specialist retains its dimension-specific role and score revisions remain bounded. The revised judgments are then carried forward as the agent states for the next round, and the same reliability-guided deliberation is repeated over these updated states. Across multi-round deliberations, the resulting corrections propagate among the participating agents while preserving dimension-specific differences supported by the available evidence. After the final round, the revised score is denoted by s'_d, while residual uncertainty in the revised judgment is retained as u_d. To stabilize the multi-round revisions, the final deliberation score is combined with its own pre-deliberation estimate:

$$\widetilde{s}_d = \gamma_d s'_d + (1 - \gamma_d) s_d, \tag{3}$$

where $\gamma_d \in [0, 1]$ controls the contribution of the revised judgment relative to the original dimension-specific estimate. This preserves the initial evidence-grounded judgment as a dimension-specific anchor while incorporating corrections introduced through deliberation. The resulting council prior $\widetilde{s}_d$, together with its supporting evidence E_d, evidence reliability ρ_d, and residual uncertainty u_d, is passed to the second-level preference optimization.

4.3 RC-MEPO: RUBRIC-CONDITIONED MULTI-EXPERT PREFERENCE OPTIMIZATION

Building on the council prior $\widetilde{s}_d$ from the first level, the resulting assessment is grounded in multimodal evidence but does not yet capture how expert raters apply the dimension-specific rubric $\mathcal{R}_d$. To align the council prior with expert judgments, we introduce RC-MEPO and optimize it in a two-stage pipeline. For dimension d, we denote the policy input as $x_d = (E_d, \widetilde{s}_d, \rho_d, u_d, \mathcal{R}_d)$. For the first stage, we use supervised fine-tuning (SFT) (Ouyang et al., 2022) to learn expert corrections around the council prior. Given the score s_{rd} assigned by expert rater r, we define the supervised target as:

$$s^*_{rd} = \widetilde{s}_d + \omega\,(s_{rd} - \widetilde{s}_d)\,, \tag{4}$$

where $s_{rd} - \widetilde{s}_d$ denotes the expert correction and $\omega \in [0, 1]$ controls its contribution. The target is rounded to the nearest valid score level before SFT, and the resulting policy is frozen as π_{ref} for the subsequent stage.

For the second stage, we perform on-policy preference optimization (PO) over the full expert panel. Since experts may assign different scores under the same rubric, we construct a confidence-weighted expert score distribution q_d from their scores and confidence ratings rather than collapsing the panel to a mean score. Starting from π_{ref}, the current policy π_θ samples candidate assessments o_i, each with a candidate rationale z_i, and optimizes the KL-regularized objective (Ouyang et al., 2022; Kirk et al., 2024):

$$\mathcal{J}(\theta) = \mathbb{E}_{x_d \sim \mathcal{D}} \left[\mathbb{E}_{o_i \sim \pi_\theta(\cdot|x_d)} R_{\text{MEPO}}(o_i; x_d, q_d) - \beta_{\text{KL}} D_{\text{KL}} \left(\pi_\theta(\cdot \mid x_d) \,\|\, \pi_{\text{ref}}(\cdot \mid x_d) \right) \right], \tag{5}$$

where β_{KL} controls regularization toward π_{ref}, and the base reward R_0 from R_{MEPO} captures pointwise score agreement. To move beyond this pointwise signal, we introduce Expert-Distribution Alignment (EDA), which compares the model's score distribution conditioned on each candidate rationale with the confidence-weighted expert score distribution q_d:

$$R_{\text{EDA}}(o_i; x_d, q_d) = -\sum_{k=1}^{5} q_d(k) \log \frac{q_d(k)}{p_\theta(k \mid x_d, z_i)}, \tag{6}$$

where $p_\theta(k \mid x_d, z_i)$ is obtained by normalizing the logits of the five score tokens after conditioning on the candidate rationale z_i. EDA thus assigns higher rewards to candidates whose score distributions better match the confidence-weighted expert distribution.

To further ensure that scoring follows the supplied rubric, we introduce the Rubric-Following Constraint (RFC). For each candidate, we keep z_i and all non-rubric components of x_d fixed and perturb only $\mathcal{R}_d$ to stricter and looser variants, obtaining x_d^- and x_d^+:

$$\mu_{\theta,i}(x_d^-) + \delta \le \mu_{\theta,i}(x_d) \le \mu_{\theta,i}(x_d^+) - \delta, \tag{7}$$

where $\mu_{\theta,i}(x)$ denotes the expected score under the rationale-conditioned score distribution for candidate o_i, and $\delta > 0$ specifies the required score margin. Violations of this constraint define the candidate-level reward $R_{\text{RFC}}(o_i; x_d)$. The resulting RC-MEPO reward is:

$$R_{\text{MEPO}}(o_i; x_d, q_d) = R_0(o_i; x_d, q_d) + \lambda_1 R_{\text{EDA}}(o_i; x_d, q_d) + \lambda_2 R_{\text{RFC}}(o_i; x_d), \tag{8}$$

where λ_1 and λ_2 weight EDA and RFC, respectively. Both EDA and RFC are evaluated as candidate-level scalar rewards and contribute to policy updates through signed group-relative advantages (Shao et al., 2024). At inference, multiple judgments are sampled, and their score distributions are averaged before computing the final score. Additional implementation details are provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Datasets and Implementation Details. For the main experiments on the MEPA dataset, we perform five-fold cross-validation over the 35 presentations, using 28 for training and 7 for held-out evaluation in each fold. The same folds are used across all configurations, with all label-dependent calibration estimated exclusively on the training fold and fixed before evaluation. For out-of-domain evaluation, we collect 12 presentations from a separate university course under the same recording protocol, independently assessed by two experienced teaching assistants and reserved for evaluation. We further evaluate on 500 stratified SpeechOcean762 (Zhang et al., 2021) test utterances, mapping its accuracy, fluency, and prosody annotations to the corresponding speech rubrics without target-domain supervision or parameter updates. RC-MEPO adapts Qwen3-8B with 4-bit QLoRA and is trained on NVIDIA L40S GPUs for two epochs per stage using AdamW with a learning rate of 1×10^{-4}. SFT uses an effective batch size of 16, while PO samples 6 candidates per instance. More implementation details are provided in Appendices B and C.

Metrics and Baselines. For in-domain assessment, we report Spearman's ρ (Spearman, 1904), mean absolute error (MAE) (Willmott & Matsuura, 2005), ordinal Krippendorff's α (Krippendorff, 2011), and the standard deviation ratio (SDR) $\sigma_{\text{model}}/\sigma_{\text{expert}}$ (Wang & von Davier, 2014) to measure ranking consistency, absolute error, ordinal agreement, and score dispersion, respectively. For out-of-domain evaluation, we additionally report Pearson's r (Pearson, 1896), Kendall's τ (Kendall

Table 1: In-domain comparison on the MEPA dataset under single-expert and five-expert consensus references. The single-expert reference uses one randomly selected expert, while the five-expert consensus is weighted by the rater confidence scores. Brackets report 95% presentation-level bootstrap confidence intervals, and † denotes proprietary models. For MoE models, Params. report total and active parameter counts. For Ours, it reports the parameter count of the RC-MEPO judge backbone. Best and second-best results are shown in **bold** and underlined.

Method	Params.	Single-Expert Reference				Five-Expert Consensus				
		$\rho \uparrow$		MAE↓	$\alpha \uparrow$	SDR$\rightarrow 1$	$\rho \uparrow$		MAE↓	$\alpha \uparrow$	SDR$\rightarrow 1$
Qwen2.5-Omni	7B	-0.007	[-0.35,0.35]	1.15	-0.24	0.34	0.073	[-0.28,0.43]	0.91	-0.37	0.53
Ola	7B	0.131	[-0.19,0.43]	0.75	0.12	0.53	0.295	[0.00,0.55]	0.59	0.15	0.82
Baichuan-Omni-1.5	7B	0.085	[-0.26,0.42]	0.93	-0.15	0.37	0.281	[-0.09,0.60]	0.63	-0.34	0.57
MiniCPM-o-2.6	8B	0.215	[-0.07,0.46]	0.71	0.18	0.47	0.147	[-0.20,0.45]	0.65	0.12	0.73
MiniCPM-o-4.5	9B	0.493	[0.15,0.74]	0.90	0.03	0.41	0.447	[0.08,0.72]	0.61	-0.09	0.63
Qwen3-Omni-Instruct	30B/3B	0.173	[-0.17,0.49]	0.89	-0.18	0.41	0.323	[0.00,0.60]	0.59	-0.26	0.64
Qwen3-Omni-Thinking	30B/3B	0.228	[-0.14,0.56]	0.76	0.28	**0.64**	0.424	[0.10,0.68]	0.55	0.49	**1.00**
Gemini 3.1 Pro†	–	0.599	[0.36,0.78]	0.62	**0.56**	0.36	0.764	[0.56,0.89]	0.39	0.58	0.55
Xiaomi MiMo-v2.5†	–	0.485	[0.17,0.74]	0.63	0.47	0.47	0.556	[0.25,0.78]	0.47	0.47	0.72
Ours	8B	**0.813**	[0.61,0.92]	**0.55**	0.50	0.62	**0.875**	[0.73,0.93]	**0.27**	**0.63**	0.76

Table 2: Out-of-domain comparison on the Cross-Course Evaluation Dataset and SpeechOcean762 without target-domain supervision or parameter updates. Brackets report 95% bootstrap confidence intervals. Bootstrap resampling is performed over presentations for the cross-course dataset and over utterances for SpeechOcean762. Best and second-best results are shown in **bold** and underlined. Dashes denote undefined correlations caused by constant predictions.

Method	Cross-Course Evaluation Dataset					SpeechOcean762			
	$r \uparrow$	$\rho \uparrow$		$\tau \uparrow$	Verbosity Bias↓	$r \uparrow$	$\rho \uparrow$		$\tau \uparrow$
Qwen2.5-Omni	0.023	-0.034	[-0.50,0.41]	-0.029	0.441	0.197	0.179	[0.12,0.23]	0.161
Ola	0.064	0.088	[-0.23,0.39]	0.076	0.187	–	–		–
Baichuan-Omni-1.5	0.162	0.190	[-0.06,0.39]	0.170	0.116	0.022	0.007	[-0.05,0.06]	0.007
MiniCPM-o-2.6	0.019	0.070	[-0.19,0.35]	0.060	0.213	0.184	0.185	[0.13,0.24]	0.165
MiniCPM-o-4.5	0.290	0.243	[-0.33,0.68]	0.261	0.621	0.221	0.226	[0.11,0.36]	0.173
Ours	**0.346**	**0.351**	[0.06,0.56]	**0.292**	**0.018**	**0.269**	**0.289**	[0.24,0.34]	**0.216**

et al., 1938), and verbosity bias (Saito et al., 2023; Zheng et al., 2023) to characterize assessment agreement and sensitivity to presentation length. We compare our method with seven open-weight omni-modal MLLMs: Qwen2.5-Omni (Xu et al., 2025), Qwen3-Omni-Instruct and Qwen3-Omni-Thinking (Team, 2025), MiniCPM-o-2.6, MiniCPM-o-4.5 (Cui et al., 2026), Ola (Liu et al., 2025), and Baichuan-Omni-1.5 (Inc., 2025), together with Gemini 3.1 Pro and Xiaomi MiMo-V2.5. All models receive matched multimodal inputs and scoring rubrics, and 95% confidence intervals are reported from 10,000 bootstrap resamples (Efron, 1979).

5.2 Comparison Analysis

In-domain analysis. Table 1 shows that our method achieves the highest presentation-level Spearman correlation and lowest MAE under both single-expert and five-expert evaluation references, together with the highest ordinal agreement against the expert consensus. Compared with the 7-9B open-weight baselines, the advantage becomes more pronounced under the panel consensus, indicating closer alignment with the shared scoring tendency. The same pattern holds against Gemini 3.1 Pro, where our method achieves higher ρ and ordinal α together with lower MAE under the consensus reference. The SDR results further separate score dispersion from presentation-level alignment. Although Qwen3-Omni-Thinking closely matches the expert score dispersion in terms of SDR, its presentation-level correlation remains substantially lower at ρ, suggesting that our stronger alignment comes from more accurate scoring across presentations rather than simply reproducing the overall spread of expert ratings.

Out-of-domain analysis. Table 2 evaluates out-of-domain performance under two complementary distribution shifts. The Cross-Course Evaluation Dataset follows the same recording protocol but changes the presentation content, presenters, and raters, while SpeechOcean762 further changes the corpus, input modality, and scoring rubric. On the cross-course dataset, our method achieves the highest agreement across all three correlation measures and reduces verbosity bias from 0.621 to 0.018 compared with MiniCPM-o-4.5, indicating that the stronger agreement is not driven by presentation length. On SpeechOcean762, our method maintains the highest Pearson and Spearman correlations despite the larger domain shift. These results indicate that the learned assessment behavior generalizes beyond the training course and remains effective across substantially different evaluation conditions.

Table 3: Component ablation against the five-expert consensus. MEE denotes Multimodal Evidence Extraction, ER denotes Evidence Reliability, and MACG denotes the Multi-Agent Collaboration Graph. RC-MEPO (SFT) denotes the supervised fine-tuning stage, while RC-MEPO (PO) denotes the subsequent preference optimization stage. EDA and RFC denote Expert-Distribution Alignment and the Rubric-Following Constraint. The fixed regressor serves only as an aggregation control. The final four rows report mean ± standard deviation over three training seeds.

Configuration	Enabled Modules							Evaluation Metrics		
	MEE	ER	MACG	SFT	PO	EDA	RFC	$\rho \uparrow$	MAE$\downarrow$	AUROC$\uparrow$
Transcript-only judge	–	–	–	–	–	–	–	0.374	0.86	0.499
Multimodal Evidence Extraction	✓	–	–	–	–	–	–	0.633	0.62	0.499
+ Evidence Reliability	✓	✓	–	–	–	–	–	0.704	0.53	0.548
+ Multi-Agent Collaboration Graph	✓	✓	✓	–	–	–	–	0.722	0.38	0.548
Fixed-Regressor Alignment	✓	✓	✓	–	–	–	–	0.623	0.32	0.548
RC-MEPO (SFT)	✓	✓	✓	✓	–	–	–	0.767±0.013	0.313±0.006	0.548
RC-MEPO (PO)	✓	✓	✓	✓	✓	–	–	0.867±0.010	0.268±0.009	0.548
+ Expert-Distribution Alignment	✓	✓	✓	✓	✓	✓	–	0.862±0.003	0.269±0.002	0.548
Ours	✓	✓	✓	✓	✓	✓	✓	**0.875**±0.016	**0.266**±0.003	0.548

Table 4: Reliability ranking across three speech-related dimensions of the MEPA dataset. Retained-set Spearman's ρ is reported at different coverage levels, with $\overline{\rho}_{\text{sel}}$ summarizing performance across the coverage range. Best and second-best results are shown in **bold** and underlined.

Reliability signal	Retained-set $\rho \uparrow$			$\overline{\rho}_{\text{sel}} \uparrow$
	100%	75%	50%	
LLM verbalized confidence	0.661	0.612	0.532	0.616
LLM score self-consistency	0.661	0.623	0.643	0.647
Audio softmax confidence	0.661	0.599	0.536	0.587
Evidence Reliability (ER)	0.661	**0.685**	**0.706**	**0.688**

Table 5: Cross-seed stability of RC-MEPO variants. Mean-dim. ρ averages Spearman correlation across the eight scoring dimensions. Values are mean ± standard deviation over three seeds. Best and second-best are shown in **bold** and underlined.

Config.	Mean-dim. $\rho \uparrow$	$\alpha \uparrow$	SDR$\rightarrow 1$
RC-MEPO	0.629±0.045	0.630±0.001	0.746±0.013
RC-MEPO + EDA	0.678±0.009	0.630±0.000	0.762±0.000
Ours	**0.682**±0.015	0.630±0.001	**0.768**±0.003

5.3 Ablation Studies

Table 3 isolates the contributions of the main components in our method. Starting from the transcript-only baseline, MEE substantially improves both ranking agreement and absolute scoring accuracy, showing the benefit of grounding assessment in multimodal evidence. Adding ER further improves expert alignment and raises the error-detection AUROC (Fawcett, 2006) from 0.499 to 0.548. MACG yields a smaller change in ρ but reduces MAE, indicating that reliability-guided collaboration mainly refines absolute score assignments while preserving presentation ordering. Meanwhile, the fixed-regressor control lowers MAE but reduces ρ, showing that closer pointwise fitting does not necessarily improve expert-aligned ranking. In contrast, RC-MEPO (SFT) improves both measures over the council prior, while RC-MEPO (PO) contributes the larger ranking gain, increasing ρ from 0.767 to 0.867 with a further reduction in MAE. This separation reflects the different roles of the two stages, where SFT learning supervised corrections around the council prior, and PO primarily improves expert-aligned ordering.

5.4 Additional Analyses

5.4.1 Reliability for Selective Retention

To assess whether evidence reliability (ER) can identify less reliable assessments, we use four reliability signals to rank the same set of held-out predictions and progressively remove the least reliable cases. For the two Qwen3-8B baselines, we run the unadapted model eight times on the extracted evidence and rubric. Specifically, LLM verbalized confidence uses the mean generated confidence, while LLM score self-consistency uses the negative standard deviation of the predicted scores. Audio softmax confidence takes the maximum probability that Qwen2.5-Omni-7B assigns to the five score levels from the original audio and rubric. As shown in Table 4, ER is the only signal whose retained-set correlation exceeds the full-coverage result at both 75% and 50% coverage and overall coverage. These results show that ER more effectively ranks assessments by reliability and supports the deferral of less reliable cases, without treating ρ_d as a calibrated probability of correctness.

5.4.2 Distributional and Rubric Alignment

Expert-Distribution Alignment. Table 5 first examines the behavior of EDA across training seeds. Adding EDA raises the mean dimension-level ρ and reduces its variation across seeds, while ordinal agreement and score dispersion remain largely unchanged. Meanwhile, we further examine whether

Table 6: Distributional alignment between model and confidence-weighted expert score distributions on held-out presentation-dimension pairs. Values are mean ± standard deviation over three seeds, and lower is better. Best and second-best are shown in **bold** and underlined.

Metric	RC-MEPO	RC-MEPO + EDA	Ours
$D_{\mathrm{KL}}(q_d \Vert \widehat{p}_d) \downarrow$	2.597±0.294	2.491±0.271	**1.744**±0.674
Brier score ↓	0.574±0.053	0.537±0.050	**0.394**±0.111
TV distance ↓	0.600±0.024	0.569±0.026	**0.509**±0.050
Spread gap ↓	0.672±0.004	**0.623**±0.011	0.656±0.014
Mean absolute error ↓	0.778±0.030	0.708±0.036	**0.668**±0.043

Table 7: Rubric-following analysis of RFC on held-out evaluation presentation samples. Results are reported as mean ± standard deviation over three seeds, and higher is better.

Metric	RFC-off	**RFC-on**
Strict consistency ↑	54.4±11.5%	**73.5**±13.3%
Loose consistency ↑	40.0±12.8%	**83.1**±3.5%
$\Delta_{\mathrm{strict}} \uparrow$	0.041±0.047	**0.121**±0.024
$\Delta_{\mathrm{loose}} \uparrow$	0.037±0.014	**0.161**±0.009

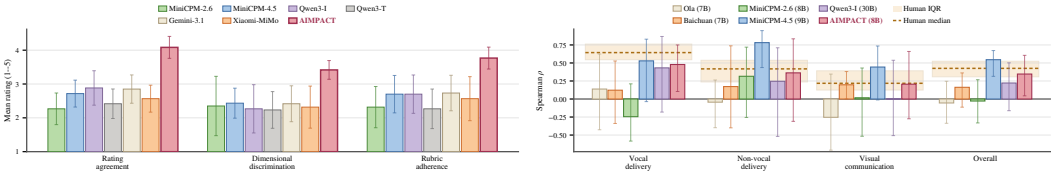

Figure 4: Human evaluation on MEPA and Cross-Course Evaluation Dataset. Left: agreement comparison with expert scores, dimensional discrimination, and rubric adherence for competing models. Right: Spearman correlation between model predictions and ratings from live audience members across vocal delivery, non-vocal delivery, visual communication, and overall performance.

this gain is reflected in the predicted score distribution. For each held-out presentation-dimension pair, we draw 32 scores to estimate the empirical predictive distribution $\widehat{p}_d$ and compare it with the confidence-weighted expert distribution q_d. In Table 6, RC-MEPO uses only the base preference reward, while adding EDA improves all five distributional metrics. Meanwhile, the pointwise ρ and MAE in Table 3 change little, indicating that EDA mainly improves how probability mass is distributed across plausible expert scores rather than shifting the final score prediction.

Rubric-Following Constraint. We compare two matched RC-MEPO variants that differ only in whether RFC is included during preference optimization. For each held-out assessment, we keep the presentation evidence and candidate rationale fixed and replace the original rubric with stricter or looser versions. Strict and loose consistency measure the proportion of cases in which the expected score changes in the required direction, while Δ_{strict} and Δ_{loose} measure the corresponding score changes. As shown in Table 7, RFC increases strict consistency from 54.4% to 73.5% and loose consistency from 40.0% to 83.1%, with larger score changes in both directions. Since the evidence and rationale are unchanged, the improvement reflects greater sensitivity to the supplied rubric, allowing RC-MEPO to adjust its scoring behavior when the evaluation standard changes.

5.4.3 Human Evaluation

To complement the automatic evaluation, we conduct two human studies that examine assessment quality and agreement with audience judgments. On MEPA, 10 evaluators rate anonymized assessments for six held-out presentations on a five-point scale. The three criteria assess agreement with expert scores, discrimination between stronger and weaker presentation dimensions, and adherence of the score and rationale to the supplied rubric. As shown in the left panel of Figure 4, AIMPACT achieves mean ratings of 4.08, 3.42, and 3.77, compared with the highest competing ratings of 2.88, 2.43, and 2.73. On the Cross-Course Evaluation Dataset, each presentation is independently rated by 12 audience members in the original classroom setting. The right panel of Figure 4 shows that AIMPACT achieves the highest overall agreement with these ratings across vocal delivery, non-vocal delivery, visual communication, and overall performance, providing an independent human reference beyond the expert annotations used for training.

6 Conclusion

We introduced AIMPACT, an evidence-grounded presentation assessment framework that combines an Evidence-to-Judgment Multi-Agent method with a Multi-Expert Presentation Assessment Dataset. Across in-domain assessment, out-of-domain assessment, and complementary human and diagnostic analyses, AIMPACT achieves the strongest expert alignment among comparable-scale MLLMs and remains competitive with substantially larger proprietary models, while generalizing to unseen presentation data and different scoring standards.

AI USE STATEMENT

Generative AI tools were used solely to assist with language editing and proofreading of the manuscript, including improvements to grammar, clarity, and academic expression. They were not used to generate scientific content, experimental results, or research conclusions. All AI-assisted revisions were carefully reviewed and verified by the authors.

REPRODUCIBILITY STATEMENT

Details required to reproduce our experiments are provided in the Appendix. Appendix A documents dataset construction and annotation, Appendix B specifies the implementation of the proposed method, and Appendix C reports the experimental protocols. The remaining sections describe the human studies, supplementary mechanism analyses, representative assessment traces, and release conditions. The code and experimental configurations will be released upon publication, while the original presentation recordings remain access-controlled due to participant privacy.

ETHICS STATEMENT

All student participants whose presentations are included in the proposed dataset were informed of the research use and provided consent to participate. Direct identifiers and sensitive information were removed or de-identified before annotation and analysis. Since the audio-visual recordings may still contain potentially identifiable information such as faces, voices, and presentation content, they are not publicly released and remain under controlled institutional access. Participants in the human evaluation studies were also informed of the study procedures, participated voluntarily, and received compensation for their participation. Expert assessment can involve subjective variation, which is why the proposed dataset retains independent judgments from five instructors rather than relying on a single rater. AIMPACT is intended to support presentation practice and formative feedback and is not designed to replace human judgment in high-stakes academic or professional decisions.

REFERENCES

Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net : Fine-grained gaze estimation in unconstrained environments. In *8th International Conference on Frontiers of Signal Processing, ICFSP 2023, Corfu, Greece, October 23-25, 2023*, pp. 98–102. IEEE, 2023.

Álvaro Becerra, Ruth Cobos, and Roberto Daza. A multimodal dataset of student oral presentations with sensors and evaluation data. *CoRR*, abs/2601.07576, 2026.

Beatrice Biancardi, Mathieu Chollet, and Chloé Clavel. Introducing the 3mt_french dataset to investigate the timing of public speaking judgements. *Lang. Resour. Evaluation*, 59(1):371–390, 2025.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Di Chang, Yufeng Yin, Zongjian Li, Minh Tran, and Mohammad Soleymani. Libreface: An open-source toolkit for deep facial expression analysis. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pp. 8190–8200. IEEE, 2024.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6562–6595, 2024.

Lei Chen and Chong Min Lee. Predicting audience's laughter during presentations using convolutional neural network. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, BEA@EMNLP 2017, Copenhagen, Denmark, September 8, 2017*, pp. 86–90. Association for Computational Linguistics, 2017.

Lei Chen, Chee Wee Leong, Gary Feng, and Chong Min Lee. Using multimodal cues to analyze mla'14 oral presentation quality corpus: Presentation delivery and slides quality. In *Proceedings of the 2014 ACM workshop on Multimodal Learning Analytics Workshop and Grand Challenge, MLA@ICMI 2014, Istanbul, Turkey, November 12, 2014*, pp. 45–52. ACM, 2014.

Lei Chen, Gary Feng, Chee Wee Leong, Jilliam Joe, Christopher Kitchen, and Chong Min Lee. Designing an automated assessment of public speaking skills using multimodal cues. *J. Learn. Anal.*, 3(2):261–281, 2016.

Luyu Chen, Zeyu Zhang, Haoran Tan, Quanyu Dai, Hao Yang, Zhenhua Dong, and Xu Chen. Distributional llm-as-a-judge. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, Jiancheng Gui, Luoyuan Zhang, Xian Sun, Fuwei Huang, Moye Chen, Zhuo Lin, Hanyu Liu, Qingxin Gui, Qingzhe Han, Yuyang Wen, Huiping Liu, Rongkang Wang, Yaqi Zhang, Hongliang Wei, Chi Chen, You Li, Kechen Fang, Jie Zhou, Yuxuan Li, Guoyang Zeng, Chaojun Xiao, Yankai Lin, Xu Han, Maosong Sun, Zhiyuan Liu, and Yuan Yao. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *CoRR*, abs/2604.27393, 2026.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric P. Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 4443–4458. Association for Computational Linguistics, 2020.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1):1 – 26, 1979. doi: 10.1214/aos/1176344552.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006. doi: 10.1016/J.PATREC.2005.10.010.

Baichuan Inc. Baichuan-omni-1.5 technical report. *CoRR*, abs/2501.15368, 2025.

Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *CoRR*, abs/2303.07399, 2023.

Maurice G Kendall et al. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024a.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 4334–4353. Association for Computational Linguistics, 2024b.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*, 2024.

Klaus Krippendorff. Computing Krippendorff's alpha-reliability. *Departmental Papers, Annenberg School for Communication, University of Pennsylvania*, 2011.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pp. 11286–11315. Association for Computational Linguistics, 2024.

Haley Lepp, Chee Wee Leong, Katrina Roohr, Michelle P. Martin-Raugh, and Vikram Ramanarayanan. Effect of modality on human and machine scoring of presentation videos. In *ICMI '20: International Conference on Multimodal Interaction, Virtual Event, The Netherlands, October 25-29, 2020*, pp. 630–634. ACM, 2020.

Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model with progressive modality alignment. *CoRR*, abs/2502.04328, 2025.

Xavier Ochoa and Heru Zhao. Openopaf: An open-source multimodal system for automated feedback for oral presentations. *Journal of Learning Analytics*, 11(3):224–248, 2024.

Xavier Ochoa, Federico Domínguez, Bruno Guamán, Ricardo Maya, Gabriel Falcones, and Jaime Castells. The RAP system: automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge, LAK 2018, Sydney, NSW, Australia, March 07-09, 2018*, pp. 360–364. ACM, 2018.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction, ICMI 2014, Istanbul, Turkey, November 12-16, 2014*, pp. 50–57. ACM, 2014.

Karl Pearson. Vii. mathematical contributions to the theory of evolution.—iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, (187):253–318, 1896.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Vikram Ramanarayanan, Chee Wee Leong, Lei Chen, Gary Feng, and David Suendermann-Oeft. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pp. 23–30. ACM, 2015.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *CoRR*, abs/2310.10076, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

Qwen Team. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025.

Su Shwe Yi Tun, Shogo Okada, Hung-Hsuan Huang, and Chee Wee Leong. Multimodal transfer learning for oral presentation assessment. *IEEE Access*, 11:84013–84026, 2023.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving llm-as-a-judge inference with the judgment distribution. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025*, pp. 23173–23199. Association for Computational Linguistics, 2025.

Zhen Wang and Alina A von Davier. Monitoring of scoring using the e-rater® automated scoring system and human raters on a writing test. *ETS Research Report Series*, 2014(1):1–21, 2014.

Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30:79–82, 2005.

Torsten Wörtwein, Mathieu Chollet, Boris Schauerte, Louis-Philippe Morency, Rainer Stiefelhagen, and Stefan Scherer. Multimodal public speaking performance assessment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, November 09 - 13, 2015*, pp. 43–50. ACM, 2015.

Zihua Xiong, Jiachen Tan, Tingting Zhang, Bin Wu, and Chunping Zheng. Prompt-guided multi-task decoupling for speech presentation skills assessment. In *IEEE International Conference on Multimedia and Expo, ICME 2025, Nantes, France, June 30 - July 4, 2025*, pp. 1–6. IEEE, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pp. 3710–3714. ISCA, 2021.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

AIMPACT: An Evidence-Grounded Multi-Agent Framework for Automated Presentation Assessment

Appendices

A Dataset Construction and Annotation 15

A.1 Data Collection and Preprocessing 15

A.2 Assessment Rubric 15

A.3 Expert Annotators 16

A.4 Annotation Procedure and Quality Control 17

A.5 Dataset Statistics 17

B Implementation Details 18

B.1 Multimodal Evidence Extraction 18

B.2 Multi-Agent Collaboration Graph 18

B.3 RC-MEPO Training and Inference 19

C Experimental Protocols 20

C.1 Evaluation Datasets 20

C.2 Baseline Implementations 21

C.3 Evaluation Metrics 21

D Human Evaluation 22

D.1 Generated-Assessment Evaluation 22

D.2 Live-Audience Evaluation 22

D.3 Evaluation Interfaces and Instructions 23

E Supplementary Experimental Results 23

E.1 Cross-Domain Reliability for Selective Retention 23

E.2 Collaboration and Robustness under Evidence Corruption 24

E.3 Evidence Faithfulness under Controlled Perturbations 24

F Scope, Reproducibility, and Method Positioning 25

F.1 Interpretation of the Empirical Evidence 25

F.2 Reproducible Artifacts and Data Access 25

A Dataset Construction and Annotation

A.1 Data Collection and Preprocessing

The Multi-Expert Presentation Assessment Dataset (MEPA) contains 35 graded presentations collected from a university computer science course under a shared assignment brief and assessment rubric. Speakers selected their own topics and prepared their own slides, preserving variation in presentation content, delivery style, and performance quality under consistent instructional and recording conditions. Each presentation lasts 3.6-6.4 minutes, yielding 2.81 hours of synchronized recordings and 380 slides in total. To support fine-grained multimodal assessment, each presentation was captured with two fixed 4K cameras and a synchronized wireless microphone: the wide view records the speaker together with the projected slides, while the close view provides clearer evidence of facial expression, gaze, and upper-body behavior. The original slide deck was retained and rendered page by page. Before annotation, all streams were aligned to a common timeline, checked for missing or corrupted segments, assigned anonymous presentation identifiers, and stripped of direct identifying information. Preprocessing then derives the modality-specific inputs required by the assessment pipeline while preserving the original recordings. Specifically, audio is resampled for word-level alignment, video is sampled for nonverbal behavior analysis, and slide pages are aligned with their corresponding presentation intervals. The original multimodal recordings remain under controlled access for internal verification.

A.2 Assessment Rubric

MEPA evaluates presentation performance across eight dimensions covering language, vocal delivery, non-vocal delivery, and visual communication. All experts use the same five-level ordinal scale from Novice (1) to Advanced (5), with dimension-specific descriptors defining the evidence required at each level. The rubric therefore provides explicit behavioral anchors rather than treating the scores as generic quality labels. The complete scoring criteria presented to the expert raters are summarized below.

- **Language.** This dimension evaluates the lexical and grammatical accuracy of the speaker's language and its appropriateness for the intended audience. (5) Advanced: Language is consistently accurate and appropriate throughout the presentation. (4) Competent: Language is mostly accurate and appropriate, with minor errors that do not affect understanding. (3) Satisfactory: Occasional inaccuracies or inappropriate expressions slightly affect understanding. (2) Developing: Frequent language problems often impede understanding. (1) Novice: Persistent inaccuracies or inappropriate language substantially impede understanding.
- **Pronunciation.** This dimension evaluates the accuracy of word-level pronunciation and the extent to which pronunciation supports intelligible speech. (5) Advanced: Pronunciation is consistently accurate. (4) Competent: Most words are pronounced accurately, with only minor errors that do not affect understanding. (3) Satisfactory: Some pronunciation errors are present and may slightly affect understanding. (2) Developing: Frequent pronunciation errors noticeably impede understanding. (1) Novice: Persistent pronunciation problems substantially interfere with meaning.
- **Volume and Pace.** This dimension evaluates whether the speaker's volume and pace are appropriate and how effectively they convey confidence and enthusiasm. (5) Advanced: Volume and pace are consistently appropriate and support a confident and engaging delivery. (4) Competent: Volume and pace are generally appropriate, with confidence and enthusiasm conveyed most of the time. (3) Satisfactory: Noticeable variations in volume or pace occur, and confidence or enthusiasm is conveyed inconsistently. (2) Developing: Frequent problems with volume or pace weaken clarity and engagement. (1) Novice: Volume and pace are consistently inappropriate and substantially affect the effectiveness of the delivery.
- **Fluency.** This dimension evaluates the continuity of speech, with particular attention to interruptions, hesitations, and pauses. (5) Advanced: Speech is fluent with almost no disruptive interruptions or pauses. (4) Competent: Speech is mostly fluent, with only occasional minor interruptions or pauses. (3) Satisfactory: Noticeable interruptions or pauses occur but do not substantially affect understanding. (2) Developing: Frequent disfluencies sometimes impede understanding. (1) Novice: Persistent disfluency or long pauses substantially disrupt understanding.
- **Non-vocal Delivery.** This dimension evaluates how effectively posture, eye contact, facial expression, and gesture support non-verbal communication and audience engagement. (5) Ad-

vanced: Non-verbal behaviors consistently support effective communication and audience engagement. (4) Competent: Most non-verbal behaviors are effective, with only occasional lapses. (3) Satisfactory: Some non-verbal behaviors support audience engagement, but their effectiveness is inconsistent. (2) Developing: Most non-verbal behaviors provide limited support for communication with the audience. (1) Novice: Non-verbal behaviors provide little or no effective engagement with the audience.

- **Slide Design.** This dimension evaluates how effectively the slide visuals support and enhance the audience's understanding of the presentation content. (5) Advanced: Visuals consistently enhance understanding of the content. (4) Competent: Visuals support understanding effectively for most of the presentation. (3) Satisfactory: Visuals are generally helpful but contain noticeable lapses that reduce their effectiveness. (2) Developing: Visuals contribute little to the audience's understanding. (1) Novice: Visuals fail to support understanding and may introduce confusion.
- **Speech–Slide Synchronization.** This dimension evaluates the synchronization between the presenter's speech and the slides displayed during the presentation. (5) Advanced: Speech and visuals remain consistently synchronized throughout the presentation. (4) Competent: Speech and visuals are synchronized most of the time. (3) Satisfactory: Synchronization is partial or inconsistent across the presentation. (2) Developing: The displayed visuals are frequently out of sync with the spoken content. (1) Novice: Speech and visuals are persistently misaligned.
- **Slide Reference.** This dimension evaluates how effectively the presenter refers to, explains, and directs the audience's attention to the visual material. (5) Advanced: The presenter consistently explains the visuals and strategically directs audience attention to relevant information. (4) Competent: The presenter generally refers to and explains the visuals effectively, with occasional missed opportunities to guide attention. (3) Satisfactory: Some visual elements are referenced and explained, while others receive insufficient attention. (2) Developing: The presenter rarely refers to or explains the visuals. (1) Novice: The presenter makes no meaningful reference to the visual material.

For each presentation and assessment dimension, every expert independently provides an integer score from 1 to 5, a five-level confidence rating, and written feedback supporting the judgment. Comments can also be attached to specific timestamps when the underlying evidence is temporally localized. The five expert assessments are retained individually rather than collapsed into a single label, preserving both panel-level consensus and rater-specific variation in scores, confidence, and feedback.

A.3 Expert Annotators

The annotation panel consists of five instructors from a university English-language communication centre, all with substantial experience in English communication, oral presentation teaching, and tertiary-level assessment. To preserve anonymity, we report only background information relevant to annotation quality. Collectively, the panel covers undergraduate and postgraduate assessment across academic, professional, scientific, and workplace communication settings.

- **R1** has approximately 13 years of experience teaching English and professional communication, including presentation training for undergraduate, graduate, and workplace contexts.
- **R2** has more than 20 years of tertiary-level experience in English language and communication, with extensive assessment of workplace presentations, project proposals, and science communication for both specialist and general audiences.
- **R3** has approximately 27 years of experience teaching English communication and oral presentation skills, including assessment at both undergraduate and postgraduate levels across research, science, current affairs, and language studies.
- **R4** has approximately 18 years of experience in English communication and oral presentation teaching, including assessment of final-year projects and presentations in computing, engineering, and science.
- **R5** has approximately 20 years of teaching experience, including about 10 years in tertiary education, with presentation assessment across undergraduate and postgraduate courses in academic and professional communication.

Before formal annotation, all experts received the complete rubric, scoring and annotation instructions, and three norming presentations representing different performance levels. Each expert first assessed the norming presentations independently across all eight dimensions. The panel then par-

ticipated in a two-hour calibration session to compare ratings, clarify rubric interpretation, and establish benchmark ratings for the three examples. These benchmarks remained available as references during formal annotation, while all study presentations were still assessed independently by each expert. Annotators were instructed to review each complete presentation before assigning dimension-level scores, confidence ratings, and written evidence. As shown in Figure 5, the annotation interface presented the video, dimension-specific rubric, score, confidence rating, and evidence field within a unified workspace.

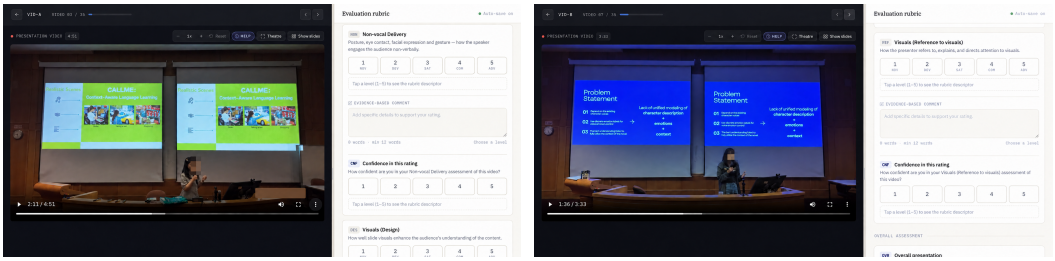

Figure 5: The expert annotation workspace for different presentation samples at non-initial timestamps. The left panel of the workspace displays the rubric and response fields for non-vocal delivery, while the right panel displays the corresponding interface for slide reference.

A.4 Annotation Procedure and Quality Control

The experts annotated the presentations independently through a browser-based platform. For each dimension, the interface displayed the relevant rubric descriptors and provided fields for the score, confidence rating, and written justification. Annotators could pause, replay, and seek within a presentation. They were instructed to base each score on the complete performance and to use timestamped comments when a specific event materially affected the judgment. The platform required all eight dimensions to be completed before a review could be submitted. Every submission was checked for missing scores, confidence ratings, and feedback. Score ranges and presentation identifiers were then validated before the records entered the analysis pipeline, and the complete five-rater matrix was inspected for accidental duplicates and incomplete reviews. No score was changed to increase agreement. The final dataset therefore contains five independent reviews for every presentation, yielding 175 complete reviews and 1,400 dimension-level judgments.

A.5 Dataset Statistics

Table 8 reports the score distribution and inter-rater reliability for each dimension. We use the two-way random-effects formulation of the intraclass correlation coefficient (Shrout & Fleiss, 1979). ICC(2,1) estimates the reliability of a single randomly selected expert, while ICC(2,k) estimates the reliability of the mean across the five experts. The latter is the relevant quantity for the five-expert consensus used in our experiments.

Table 8: Per-dimension annotation statistics for the MEPA dataset. ICC(2,1) measures single-rater reliability, ICC(2,k) measures the reliability of the five-rater mean, and pairwise r is the mean Pearson correlation across rater pairs.

Dimension	Mean	Std.	ICC(2,1)	ICC(2,k)	Pairwise r
Language	3.76	0.66	0.504	0.836	0.573
Pronunciation	3.40	0.69	0.479	0.821	0.573
Volume and pace	3.44	0.70	0.613	0.888	0.678
Fluency	3.57	0.64	0.577	0.872	0.625
Non-vocal delivery	3.03	1.06	0.717	0.927	0.794
Slide design	3.52	0.53	0.380	0.754	0.396
Speech–slide synchronization	3.95	0.58	0.443	0.799	0.553
Slide reference	3.21	0.76	0.377	0.752	0.524

Single-rater reliability varies across dimensions, with ICC(2,1) ranging from 0.377 to 0.717, while averaging the five ratings raises ICC(2,k) to between 0.752 and 0.927. The mean expert confidence is 3.96 with a standard deviation of 0.51, and 86% of the judgments receive a confidence rating of at least 4. The dimensions remain related but not interchangeable. Their mean absolute off-diagonal correlation is 0.695, and the first two principal components explain 83.6% of the variance. These statistics support a panel reference while retaining dimension-specific and rater-level variation.

B Implementation Details

B.1 Multimodal Evidence Extraction

Multimodal Evidence Extraction (MEE) converts the raw audio, video, and slide streams into modality-specific evidence records before any rubric-level judgment is made. Audio is resampled to 16 kHz and transcribed with word-level timestamps using 30-second decoding chunks and a beam width of 5. Voice-activity filtering is disabled to preserve short hesitation tokens for fluency analysis, and inter-word gaps of at least 0.3 seconds are treated as pauses. From the resulting transcript, speech operators extract evidence related to language, pronunciation, fluency, speaking rate, pauses, hesitations, fillers, and prosodic variation. Speech and visual delivery are summarized over non-overlapping 30-second windows, with a shorter final window when less than 30 seconds remain. Video operators additionally sample one frame per second to characterize gaze, pose, gesture, and facial activity. Slide analysis captures layout, text and visual density, as well as the temporal relation between spoken content and displayed material. Each observation is aligned to its corresponding presentation interval or slide page and stored as a structured evidence record containing the measured value, temporal or slide location, number of supporting observations, operator confidence, and provenance. These records remain descriptive measurements rather than presentation scores. Missing evidence is represented explicitly, while observations that fail an operator's validity criteria are excluded rather than assigned numerical values. The resulting evidence records are then routed to the seven specialists following the taxonomy in Section 4.2.2.

B.2 Multi-Agent Collaboration Graph

Each specialist first evaluates its assigned evidence independently and outputs a dimension score, an evidence-grounded rationale, and the identifiers of the supporting records. Evidence Reliability (ER) measures the quality of this support by combining the amount of valid evidence with the confidence provided by the corresponding operator. For count-based evidence, support is mapped through $1 - \exp(-n/\tau)$, with $\tau = 10$ sampled frames for video and $\tau = 8$ recognized words for speech. Other reliability terms retain their operator-specific definitions. Prosody uses voiced-frame coverage, pronunciation combines transcript coverage with word-level recognition confidence, and non-vocal delivery combines valid-frame coverage with keypoint completeness. Deterministic slide records receive unit confidence when present. All reliability values are normalized to $[0, 1]$ before routing, but are not interpreted as calibrated probabilities across operators.

Evidence cells with reliability below 0.3 are masked before specialist scoring. For dimension d, ρ_d is computed as the mean reliability of the nonzero evidence cells consumed by its specialist, and the three most reliable cells are retained as supporting citations. Missing cells remain explicit in the evidence memory with zero reliability but do not enter the mean when valid evidence is available. If no valid cell remains, ρ_d is set to zero. Thus, ρ_d reflects the quality of the evidence available to a specialist rather than the proportion of the presentation covered by valid observations, and serves as a routing signal rather than a calibrated estimate of prediction uncertainty.

MACG operates within groups of specialists that rely on the same input modality. Collaboration is activated only when the group has a score range of at least 1.0, a reliability range of at least 0.15, and a maximum reliability of at least 0.5. A score difference alone is therefore insufficient to trigger collaboration, and differences between similarly supported dimensions are preserved rather than treated as conflicts. Once the gate is activated, the specialists deliberate over their current scores, reliability values, and evidence-grounded rationales. Judgments supported by more reliable evidence provide stronger references during revision, while each specialist remains responsible for its assigned rubric dimension and cannot introduce evidence beyond the supplied rationales. Scores remain within $[1, 5]$ and may change by at most one point per round. Deliberation terminates when the scores stabilize or after two rounds, and the collaboration trace records the inputs and score revisions for each activated group.

After deliberation, residual uncertainty reflects both the evidence reliability of dimension d and its remaining deviation from reliable same-modality peers. Let s'_d denote the final revised score, let P_d contain peers with reliability of at least 0.5, and let $\nu(s) = (s - 1)/4$ map scores to $[0, 1]$. We compute

$$u_d = \mathrm{clip}_{[0,1]}\left(0.5(1 - \rho_d) + 0.5\left|\nu(s'_d) - \frac{1}{|P_d|}\sum_{j \in P_d}\nu(s'_j)\right|\right). \tag{9}$$

The disagreement term is set to zero when P_d is empty. The resulting council trace retains both the initial specialist score and the final revised score. RC-MEPO then receives the evidence-grounded council prior $\widetilde{s}_d$ defined in Equation 3, together with ρ_d, u_d, the cited evidence, and the dimension rubric.

B.3 RC-MEPO Training and Inference

RC-MEPO uses Qwen3-8B as the judge backbone. The backbone is frozen and loaded with 4-bit NF4 double quantization, with computation performed in bfloat16. We train bias-free LoRA adapters with rank 16, scaling factor 32, and dropout 0.05 on the query, key, value, output, gate, up, and down projections. Both training stages run for two epochs on NVIDIA L40S GPUs using AdamW with a learning rate of 1×10^{-4} and weight decay of 0.01.

The supervised fine-tuning (SFT) stage uses a micro-batch size of 2 and eight gradient-accumulation steps, giving an effective batch size of 16. We use cosine learning-rate decay with a 10% warm-up. Each training instance contains the dimension-specific evidence, evidence-grounded council prior, ER, residual uncertainty, and the corresponding assessment rubric. The residual target uses $\omega = 0.3$, so the target retains 70% of the council prior and incorporates 30% of the expert correction. Supervision is applied only to completion tokens. For training instance i, let C_i denote its completion tokens and $c_i \in \{1, \ldots, 5\}$ the confidence assigned by the expert. The confidence-weighted objective is

$$\mathcal{L}_{\mathrm{SFT}} = \frac{\sum_i (c_i/5)\, |C_i|^{-1} \sum_{t \in C_i} -\log p_\theta(y_t \mid y_{<t})}{\sum_i c_i/5}. \tag{10}$$

where y_t is the target token at position t, and p_θ is the conditional probability produced by the adapter. Rubric paraphrases are sampled during SFT to expose the model to semantically equivalent formulations of the same scoring criteria.

The preference optimization (PO) stage regenerates six candidate assessments from the current policy for each presentation-dimension pair at every epoch, using temperature 1.0, top-p 0.95, and a maximum generation length of 96 tokens. The target distribution is constructed from the five retained expert ratings. Let $s_{rd} \in \{1, \ldots, 5\}$ denote the score assigned by expert r to dimension d, and let $c_{rd} \in \{1, \ldots, 5\}$ denote the associated confidence. Their confidence-weighted empirical histogram is

$$h_d(k) = \frac{\sum_r c_{rd}\, \mathbf{1}[s_{rd} = k]}{\sum_r c_{rd}}, \qquad k \in \{1, \ldots, 5\}. \tag{11}$$

The indicator $\mathbf{1}[s_{rd} = k]$ selects the score level assigned by expert r. Confidence serves only as a relative weight within the five-member panel and is not interpreted as a calibrated probability. To preserve disagreement under the small panel size, we compute the confidence-weighted score standard deviation σ_d, set $a_d = \min(0.5, 0.6\sigma_d)$, convolve h_d with the kernel $[a_d, 1, a_d]$, and renormalize the result to obtain q_d. The smoothing transfers probability mass only to adjacent ordinal levels. When all experts assign the same score, $\sigma_d = 0$ and the distribution remains concentrated at that score.

The base reward R_0 measures pointwise agreement with the expert panel using both the sampled score and the expected score of the rationale-conditioned distribution. For candidate o_i, the discrete term rewards its sampled score $\widehat{s}_i$ according to the expert-distribution mass $q_d(\widehat{s}_i)$. The continuous term penalizes the normalized absolute deviation between $\mu_{\theta,i}(x_d)$ and the confidence-weighted expert mean $\bar{s}_d$. Both terms are scaled by the mean expert confidence $\bar{c}_d/5$, and the continuous term receives a weight of 1.5. Invalid scores receive no discrete agreement reward. This keeps R_0 focused on pointwise score agreement, while EDA and RFC provide the distributional and rubric-conditioned rewards defined in Section 4.3.

For EDA, the candidate rationale z_i is teacher-forced immediately before the score field, while the sampled score itself is excluded from the prefix. The next-token logits associated with the five score tokens are normalized to obtain $p_\theta(k \mid x_d, z_i)$. This auxiliary forward pass is performed without gradient tracking, and the resulting scalar reward is detached from the computation graph. Policy gradients therefore enter only through the log probability of the sampled completion in the group-relative objective.

RFC evaluates the same candidate rationale and evidence under stricter and more lenient versions of the rubric. The strict variant instructs the model to penalize minor deficiencies and reserve high scores for nearly flawless performance. The lenient variant gives greater credit to effort and intent while discounting minor deficiencies. The dimension criterion and its five-level score anchors remain unchanged. Writing $m = -1$ for the strict variant and $m = 1$ for the lenient variant, the

implemented reward is

$$R_{\mathrm{RFC}} = -\frac{1}{2} \sum_{m \in \{-1,1\}} \max(0, \delta - m\,[\mu_{\theta,i}(x_d^m) - \mu_{\theta,i}(x_d)])\,. \tag{12}$$

The input x_d^m differs from x_d only in the rubric instruction, and $\delta = 0.3$ specifies the required shift on the five-point scale. We set the EDA and RFC weights in Equation 8 to 0.5. These rubric variants are used only to test a controlled monotonic response to changes in rubric strictness and are not treated as independently authored assessment standards. The matched RFC-off/on comparison in Table 7 isolates the effect of RFC by keeping the presentation evidence and candidate rationale unchanged.

Within each candidate group G, the complete detached rewards are normalized into signed relative advantages:

$$A_i = \mathrm{clip}\left(\frac{R_i - \mathrm{mean}_{j \in G}\, R_j}{\mathrm{std}_{j \in G}\, R_j + 10^{-4}}, -3, 3\right), \tag{13}$$

where R_i is the reward assigned to candidate i. Groups with reward standard deviation below 10^{-6} are excluded from optimization. We clip the gradient norm to 1.0 and apply a token-level KL penalty of 0.05 relative to the frozen SFT adapter.

At inference, we sample eight judgments with temperature 0.7, top-p 0.95, and a maximum generation length of 96 tokens. Their score distributions are averaged before the final expected score is computed. Each judgment follows a fixed output schema containing an evidence-grounded rationale, an integer score, and a confidence value. Dimension scores are finalized and stored before report generation so that assessment remains separate from downstream feedback generation.

C Experimental Protocols

C.1 Evaluation Datasets

We evaluate AIMPACT under three settings that cover distinct, bounded distribution shifts. MEPA measures in-domain performance under the presentation setting used for training. The Cross-Course Evaluation Dataset changes the speakers, topics, course, and raters while retaining the multimodal presentation setting, and SpeechOcean762 changes the speech corpus and the three applicable speech criteria. No labels from either out-of-domain dataset are used for training, model selection, or parameter tuning.

- **MEPA.** The dataset contains 35 multimodal presentations, each independently assessed by five experts across eight rubric dimensions. We use a fixed five-fold protocol with 28 presentations for training and 7 for testing in each fold. The same folds are used for AIMPACT, all ablations, and all training seeds. The presentation, rather than an individual rater-dimension judgment, is the independent sampling unit.
- **Cross-Course Evaluation Dataset.** The dataset contains 12 multimodal presentations collected from a separate course. Each presentation is evaluated by two teaching assistants and 12 live audience members. It is used exclusively for zero-shot evaluation and examines transfer to new speakers, topics, a new course setting, and new raters without target-domain adaptation. Given the number of presentations, the resulting correlations are treated as exploratory estimates and are reported with presentation-level intervals.
- **SpeechOcean762.** We use a fixed test subset of 500 audio utterances with accuracy, fluency, and prosody annotations. This setting removes the visual modalities and introduces a different corpus and scoring standard. It evaluates zero-shot transfer through the speech pathway only and does not test the five remaining MEPA dimensions.

For MEPA, all label-dependent statistics are estimated using only the 28 training presentations in each fold and are fixed before evaluation on the seven held-out presentations. Predictions from the five held-out folds are then concatenated to compute corpus-level metrics. The 1,400 rater-dimension judgments describe the annotation density and are not counted as 1,400 independent evaluation samples. For the Cross-Course Evaluation Dataset, the two teaching-assistant ratings serve as the primary expert reference, while the live audience ratings are analyzed separately in Appendix D. For SpeechOcean762, accuracy, fluency, and prosody are aligned with the pronunciation, fluency, and volume-and-pace dimensions of our rubric, respectively. The model receives the target rubric at inference but no target-domain labels or parameter updates. Accordingly, the two out-of-domain studies support transfer under the reported course and speech shifts; they do

not by themselves establish invariance to arbitrary courses, recording conditions, or independently designed multimodal rubrics.

C.2 Baseline Implementations

The comparison covers open-weight omni-modal models near the size of the RC-MEPO backbone, larger mixture-of-experts models, and proprietary systems. We evaluate Qwen2.5-Omni, Qwen3-Omni-Instruct, Qwen3-Omni-Thinking, MiniCPM-o-2.6, MiniCPM-o-4.5, Ola, Baichuan-Omni-1.5, Gemini 3.1 Pro, and Xiaomi MiMo-V2.5 through their official checkpoints or service interfaces without task-specific updates. For MEPA, each compatible baseline receives the same presentation content, dimension rubric, and five-level output constraint. Audio and video use the model's native multimodal interface, and slide pages are supplied in temporal order. Model-specific responses are parsed into a common schema, and failed or non-numeric responses are retried with the same prompt and decoding settings. The out-of-domain comparison in Table 2 reports the five open-weight baselines for which complete predictions were obtained on both evaluation datasets under the fixed protocol. SpeechOcean762 uses the three mapped speech rubrics, and the Cross-Course Evaluation Dataset retains the corresponding five-dimension presentation rubric. Constant predictions have undefined rank correlation and are reported as a dash rather than zero.

Table 1 is an end-to-end deployment comparison under matched evaluation inputs, not an equal-supervision comparison. AIMPACT is adapted on the 28 MEPA training presentations in each fold and uses the Evidence-to-Judgment Multi-Agent Framework (EJMA), whereas the general-purpose baselines remain frozen and receive no MEPA labels. The table therefore measures the performance of a task-adapted assessment system against zero-shot generalist models. Component attribution instead combines the cumulative comparison in Table 3 with the targeted analyses in Section 5.4. Under a fixed data split and evaluation reference, Table 3 introduces MEE, ER, MACG, RC-MEPO (SFT), RC-MEPO (PO), and EDA before reporting the complete AIMPACT configuration, rather than an isolated RFC intervention, in its final row. Fixed-Regressor Alignment provides the corresponding non-RL aggregation control. RFC is attributed through the matched RFC-off/on comparison in Table 7, which changes the constraint while holding the presentation evidence and candidate rationale fixed.

C.3 Evaluation Metrics

In-domain evaluation uses a randomly selected single-expert reference and the confidence-weighted five-expert consensus. Spearman's ρ measures rank agreement, MAE measures absolute score deviation, ordinal Krippendorff's α measures agreement on the ordered score scale, and the standard deviation ratio (SDR) compares model and expert score dispersion. Out-of-domain evaluation reports Pearson's r, Spearman's ρ, and Kendall's τ. Evaluation on the Cross-Course Evaluation Dataset also reports verbosity bias, defined as the absolute correlation between a model's presentation score and transcript length. Confidence intervals are obtained from 10,000 percentile bootstrap resamples, using presentations as the unit for MEPA and the Cross-Course Evaluation Dataset and utterances for SpeechOcean762. All dimensions and ratings belonging to a resampled item remain grouped. Trainable RC-MEPO variants are repeated with three PO seeds while data splits, SFT adapters, and evaluation prompts remain fixed.

For the distributional analysis in Table 6, $D_{\mathrm{KL}}(q_d \| \widehat{p}_d)$ compares the expert target distribution with the empirical model distribution estimated from 32 held-out samples. The Brier score is the mean squared error across the five score probabilities, total variation is half the ℓ_1 distance, and the spread gap is the absolute difference between the two score-distribution standard deviations. Distributional MAE compares their expected scores. The reported standard deviations in Tables 5 and 6 quantify variation over the three training seeds and should not be read as presentation-level confidence intervals. Selective retention ranks instances by a confidence or reliability signal and recomputes correlation after withholding the lowest-ranked 25% and 50%. It tests whether ER orders cases for deferral rather than whether ER is a calibrated probability of correctness. Since the signals in Table 4 come from different model interfaces, this analysis compares operational deferral signals and does not establish calibration superiority under a shared confidence head.

D HUMAN EVALUATION

D.1 GENERATED-ASSESSMENT EVALUATION

This study evaluates whether the generated outputs constitute meaningful presentation assessments beyond numerical agreement with expert scores. Ten compensated evaluators reviewed anonymized assessments from seven systems for six held-out MEPA presentations. Every evaluator completed all three criteria for every presentation-system pair, yielding 126 ratings per evaluator and 1,260 ratings in total. Each item included the presentation material, assessment rubric, expert reference, and one candidate assessment. Model identities and sizes were concealed, and candidate outputs were presented in randomized order. **Agreement with expert scores** measures consistency with the displayed reference. **Dimensional discrimination** measures whether the assessment distinguishes stronger and weaker aspects of the presentation. **Rubric adherence** measures whether the score and explanation address the supplied criterion. Evaluators were instructed to judge the assessment as a whole rather than favoring a response for length or verbal fluency.

The bars in the left panel of Figure 4 are averages over evaluators and presentations. Its 95% intervals use the evaluator as the independent unit: the six presentation ratings are first averaged within each evaluator, followed by a two-sided t interval over the ten evaluator means. Table 9 reports paired comparisons with the strongest competing mean for each criterion. The interval is computed over the ten evaluator-level paired differences. Exact two-sided sign-flip tests are performed for AIMPACT against all six competitors on all three criteria, with Holm correction across the resulting 18 comparisons. Individual agreement across the 126 presentation-system-criterion items is modest, with ICC(2,1) of 0.251, whereas the aggregated ten-evaluator reference reaches ICC(2,10) of 0.770. This distinction supports reporting evaluator-clustered uncertainty rather than treating the 1,260 ratings as independent. Since the expert assessment is visible, this study measures reference-conditioned assessment fidelity rather than an independent replication of the expert judgment. The live-audience study below provides the independent human reference.

Table 9: Generated-assessment evaluation with evaluator-level uncertainty. The strongest comparator is selected by the mean for each criterion. Intervals are 95% paired t intervals, and p_{Holm} corrects all 18 AIMPACT-competitor sign-flip tests.

Criterion	AIMPACT	Strongest comparator	Paired difference	p_{Holm}
Agreement with expert scores	4.08	Qwen3-Omni-Instruct (2.88)	1.20 [0.64, 1.76]	0.047
Dimensional discrimination	3.42	MiniCPM-o-4.5 (2.43)	0.98 [0.50, 1.47]	0.047
Rubric adherence	3.77	Gemini 3.1 Pro (2.73)	1.03 [0.49, 1.57]	0.047

D.2 LIVE-AUDIENCE EVALUATION

This study evaluates model alignment with judgments collected from the live presentation setting. The cohort contains 13 student presenters, and every student other than the current presenter submitted an evaluation after each talk. We retain the 12 presentations with complete multimodal records and model predictions, yielding 12 live audience ratings for each presentation. The two teaching-assistant ratings used as the expert reference in Table 2 are excluded from this audience analysis. The audience rubric covers fluency, prosody, non-vocal delivery, visual design, and slide content, with slide content defined as the average of speech-slide synchronization and slide-reference ratings. Audience scores are averaged at the presentation-dimension level, and model alignment is measured using Spearman correlation with these means. Figure 4 groups the five dimensions into vocal delivery, non-vocal delivery, visual communication, and overall performance. Model error bars are 95% percentile intervals from 10,000 bootstrap resamples of the 12 presentations, with all dimension ratings for a presentation kept together.

To characterize the consistency of the live audience, we compute leave-one-out human agreement without allowing a student's rating to contribute to that student's reference. Each eligible rater is compared with the mean of the other 11 audience members, and correlations are averaged across the evaluated dimensions. The resulting 12 agreement values have a median of 0.425, an interquartile range of 0.309-0.524, and an overall range of 0.159–0.613. The reference band in Figure 4 represents this between-rater variation. AIMPACT's overall agreement lies within the human interquartile range and is the strongest among the evaluated systems with at most 8B parameters. None of these audience members participated in MEPA annotation or RC-MEPO training, so the study provides an independent human reference from a different course and rater group. At the same time, the 12 presentation clusters yield wide intervals for several systems and categories, including the larger

models. We therefore use this analysis as descriptive evidence of cross-course audience alignment rather than a confirmatory test of model superiority.

D.3 Evaluation Interfaces and Instructions

Interfaces. Both studies use browser-based interfaces that present one evaluation item at a time. In the generated-assessment study, the presentation material, expert reference, and candidate assessment are displayed in fixed regions with identical layouts across systems. Model identities are replaced with anonymous labels to avoid revealing the source or scale of the evaluated system. In the live-audience study, the rubric is presented immediately after each talk, and the interface prevents presenters from evaluating their own presentations.

Instructions. Before evaluation, participants receive instructions describing the study objective, the meaning of each response field, and how the supplied rubric should be applied. They are asked to base their ratings on the observed presentation and the stated assessment criterion rather than topic familiarity or stylistic preference. Responses can be reviewed before submission, but submitted evaluations remain independent and are not visible to other participants. Screenshots of the evaluation interfaces and the complete instructions are provided in the supplementary material, with account credentials and other identifying information removed.

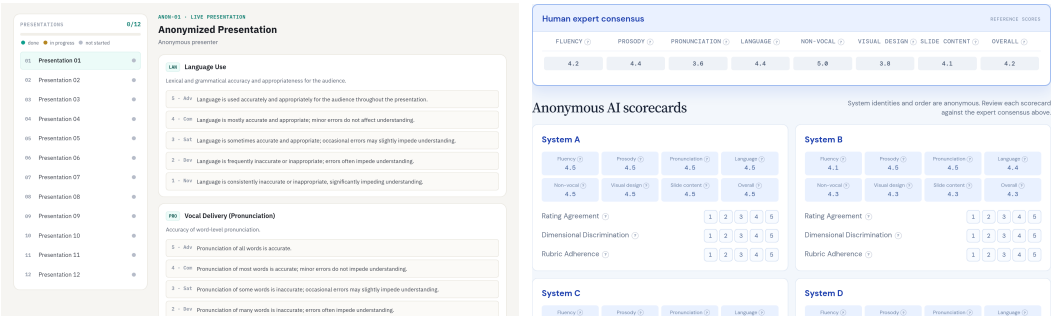

Figure 6: Interfaces used in the two human evaluations. The left panel shows the rubric-based form completed by live audience members after each presentation. The right panel shows the blinded comparison of generated assessments against the expert reference.

E Supplementary Experimental Results

The main paper evaluates AIMPACT in terms of expert alignment, distributional agreement, rubric consistency, and human judgment. This section provides additional analyses of the mechanisms underlying these results, with particular attention to reliability under distribution shift and the behavior of individual framework components.

E.1 Cross-Domain Reliability for Selective Retention

We further examine whether Evidence Reliability (ER) remains useful after transferring to SpeechOcean762, where both the data distribution and scoring rubric differ from MEPA. Table 10 first measures the association between each confidence signal and prediction accuracy using Spearman correlation with negative absolute error. Higher values therefore indicate that more confident predictions are also more accurate. ER shows the strongest positive association with correctness ($\rho_{\text{err}} = 0.107$, $p < 0.001$). LLM score self-consistency exhibits a weaker association, while LLM verbalized confidence and audio softmax confidence remain close to zero.

Table 10: Reliability for selective retention on SpeechOcean762. Left: association between each confidence signal and negative absolute error. Right: Pearson correlation after retaining the most confident predictions.

(a) Confidence and correctness

Signal	$\rho_{\text{err}} \uparrow$	p
LLM score self-consistency	0.079	0.002
LLM verbalized confidence	−0.003	0.902
Audio softmax confidence	−0.000	0.997
Evidence Reliability	**0.107**	**< 0.001**

(b) Retained-set correlation

Signal	100%	75%	50%
LLM verbalized confidence	0.253	0.231	0.237
Audio softmax confidence	0.197	0.224	0.237
Evidence Reliability	**0.289**	**0.348**	**0.404**

We then evaluate whether these confidence signals can identify predictions that should be deferred. Predictions are ranked by confidence, and the least confident cases are progressively removed before recomputing agreement on the retained subset. Because this experiment validates the Evidence Reliability signal used in council prior construction, it pairs ER with the zero-shot prediction at the council-prior stage rather than the final RC-MEPO output reported in Table 2. ER is the only signal that improves consistently as coverage decreases, with Pearson correlation increasing from 0.289 at full coverage to 0.348 at 75% coverage and 0.404 at 50% coverage. In 10,000 utterance-cluster bootstrap resamples that keep the three dimensions of each utterance together, the corresponding 95% intervals are $[0.246, 0.332]$, $[0.285, 0.400]$, and $[0.332, 0.468]$. The 50%-coverage gain over full coverage is 0.115 with an interval of $[0.067, 0.158]$. In contrast, the two retained-set baseline signals remain below 0.24 at 50% coverage. Together with the in-domain results in Table 4, these findings show that ER provides a useful ordering signal for selective retention after the reported speech-domain shift. The analysis is limited to three speech dimensions, and the comparator signals come from different model interfaces. It therefore does not establish eight-dimension calibration or a matched-head advantage over alternative confidence estimators.

E.2 Collaboration and Robustness under Evidence Corruption

Table 11: Collaboration and corruption robustness. Left: MACG on three MEPA speech dimensions with clean evidence. Right: score changes after injecting one low-reliability window per speech stream.

(a) Clean-evidence collaboration

Configuration	$r \uparrow$	$\rho \uparrow$	$\tau \uparrow$
Specialist agents	0.609	0.593	0.435
Specialist agents + MACG	**0.659**	**0.641**	**0.478**

(b) Robustness to evidence corruption

Configuration	Mean $\lvert\Delta\rvert \downarrow$	$\lvert\Delta\rvert < 0.05 \uparrow$
Without ER/MACG	0.531	7%
ER + MACG	**0.043**	**73%**

We first isolate the effect of MACG from RC-MEPO on the three speech-related MEPA dimensions. Both settings use the same extracted evidence and the same initial specialist judgments, differing only in whether reliability-guided collaboration is applied before forming the evidence-grounded council prior. As shown in the left panel of Table 11, MACG improves all three correlation measures, with Spearman correlation increasing from 0.593 to 0.641. The paired presentation-level bootstrap estimate of this Spearman change is 0.048 with a 95% interval of $[-0.009, 0.110]$. The direction is consistent with the result in Table 3, but the interval does not exclude zero and the analysis covers three speech dimensions. We next examine whether the framework remains stable when weak evidence is introduced explicitly. For 150 SpeechOcean762 utterances, one low-reliability corrupted window is injected into each of the three speech evidence streams, yielding 450 paired assessments. The added window is placed after the observed utterance, contains no recognized words, and carries deliberately adverse fluency, pronunciation, and prosody measurements. Its negligible coverage makes the intended reliability decision independent of a target-domain label. Without ER or MACG, the predicted score changes by 0.531 points on average. With ER-guided MACG, the mean change drops to 0.043, and 73% of assessments remain within 0.05 score points of the original prediction, compared with 7% without ER or MACG. The reduction in mean score displacement is 0.488, with a 95% bootstrap interval of $[0.461, 0.516]$. This controlled study tests the intended response to one weakly supported evidence record rather than the full range of naturally occurring corruption.

E.3 Evidence Faithfulness under Controlled Perturbations

Traceable evidence is meaningful only if a specialist judgment depends on the records it cites. Following the comprehensiveness and sufficiency tests for rationale faithfulness (DeYoung et al., 2020), we perturb the evidence supplied to each specialist in three controlled ways. For each specialist view, we remove the cited records, retain only the cited records, or remove an equal number of randomly selected non-cited records. Comprehensiveness measures the normalized score change after cited evidence is removed, while sufficiency measures the change when only cited evidence is retained. The faithfulness gap compares cited-evidence removal against the matched random-removal baseline, separating dependence on the cited records from sensitivity to evidence removal in general. Citation precision further measures the proportion of cited records whose removal changes the specialist score. These metrics isolate the citation-to-score link before deliberation; they do not test the linguistic plausibility of the final RC-MEPO rationale.

Table 12: Evidence faithfulness under controlled perturbations. Higher comprehensiveness, faithfulness gap, and citation precision indicate stronger dependence on the cited evidence, while lower sufficiency indicates that the cited subset better preserves the original assessment.

Specialist view	Comprehensiveness ↑	Sufficiency ↓	Faithfulness gap ↑	Citation precision ↑
Fluency	0.300	0.000	0.300	1.000
Prosody (volume and pace)	0.155	0.000	0.155	1.000
Pronunciation	0.275	0.000	0.275	1.000
Non-vocal delivery	0.128	0.165	0.128	0.667
Slide design	0.020	0.000	0.020	1.000
Slide content	0.413	0.000	0.413	1.000
Macro average	**0.215**	**0.028**	**0.215**	**0.944**

Table 12 shows positive comprehensiveness across all six evaluated specialist views, with a macro average of 0.215, indicating that removing cited evidence consistently changes the corresponding specialist score. The low average sufficiency of 0.028 shows that retaining the cited records alone largely preserves that score, while the citation precision of 0.944 indicates that most cited records have a measurable effect when removed. Language is not included because the selected controlled instance showed no response to removal of its cited records. The slide content row denotes the shared specialist evidence view used for the speech-slide synchronization and slide-reference rubrics, rather than an additional assessment dimension. Taken together, these results show that the citations used in council prior construction are functionally connected to specialist scoring. They complement the preceding analyses from a different perspective: ER evaluates whether unreliable assessments can be identified, the corruption study tests whether weak evidence is prevented from dominating collaboration, and the present experiment tests whether the evidence cited by a specialist contributes to its initial judgment.

F Scope, Reproducibility, and Method Positioning

F.1 Interpretation of the Empirical Evidence

Comparison scope. Table 1 evaluates complete assessment systems under matched presentation inputs and scoring rubrics. AIMPACT is trained on the MEPA training folds, while the general-purpose MLLMs are evaluated using their official checkpoints without task-specific adaptation. The table therefore provides an end-to-end comparison of presentation assessment systems under a common evaluation protocol. Component-level effects are examined separately in Table 3, where the data split and evaluation reference are fixed while MEE, ER, MACG, RC-MEPO (SFT), RC-MEPO (PO), and EDA are introduced progressively. Table 7 further isolates RFC through a matched RFC-off/on comparison, while the Transcript-only judge and Fixed-Regressor controls examine the contributions of multimodal evidence and learned preference alignment, respectively.

Generalization scope. The primary in-domain evaluation is conducted at the presentation level on MEPA. We further assess transfer in two complementary settings. The Cross-Course Evaluation Dataset changes the course, presenters, topics, and raters while preserving the recording protocol, whereas SpeechOcean762 introduces a broader speech-domain shift for pronunciation, fluency, and volume and pace. Together, these evaluations examine whether the learned assessment behavior transfers across changes in presentation context, annotators, corpus, and scoring conditions. Bootstrap resampling is performed at the presentation or utterance level to match the independent unit of each evaluation set.

F.2 Reproducible Artifacts and Data Access

Upon publication, we will release the implementation code and the redistributable portion of the MEPA dataset, including de-identified annotations and processed evidence records, to support reproduction of the reported experiments. The original audiovisual recordings will not be publicly released because they contain potentially identifiable faces, voices, slides, and spoken content. Access to these protected recordings will follow the applicable participant-consent and institutional requirements.