PhoenixNest-Video: Evidence-Grounded Multimodal Agent Framework for Automated Video Interview Assessment

Yuxuan Fan **Miaojun Huang** **Haimei Zhang** **Jingshen Wu** **Hao Liu**
The Hong Kong University of Science and Technology (Guangzhou)

Abstract

Interview assessment requires per-criterion judgments grounded in behavioral evidence, yet surging applicant volumes have made human-only evaluation costly and inconsistent, while existing AI approaches yield opaque scores without traceable rationale. We introduce PhoenixNest-Video, an evidence-grounded multimodal agent framework for automated video interview assessment. It builds a semantic video graph as structured working memory, performs rubric-conditioned retrieval with cross-modal verification across visual, audio, and textual streams, and produces per-criterion scores anchored to the candidate's materials. A Scorer trained via Rubrics-based Reinforcement Learning with dual rewards for rubric alignment and score-level differentiation internalizes the discriminative structure of multi-level rubrics. PhoenixNest-Video attains 91.50% grade-level accuracy on VInterview-2025, outperforming substantially larger proprietary models. A compact, rubric-grounded agent therefore scores candidates in closer agreement with an expert panel than direct prompting of much larger models, and exposes the evidence behind each score for human review.

1 Introduction

Interviews have long served as a core step in selecting candidates across graduate admissions and professional hiring (Wiens, 1976; Franklin, 2024; Albaroudi et al., 2024). To reach fair and defensible decisions, institutions require evaluators to score each candidate along multiple criteria, justify every score, and cite specific moments and materials from the interview as supporting evidence (Maude and Kirby, 2022). Yet application volumes have surged sharply in recent years, with Common Application submissions rising over 37% between 2021 and 2025 to more than 7.6 million (Song et al., 2026; Altbach et al., 2019), placing unprecedented pressure on this evidence-anchored protocol. Sustaining such rigorous per-candidate assessment with human experts alone has become increasingly impractical at this scale. On one hand, recruiting, training, and compensating qualified interviewers across thousands of applicants imposes a substantial labor cost, and expert staffing has not kept pace with demand. On the other hand, even when sufficient raters are available, prolonged large-scale evaluation amplifies well-documented human limitations such as fatigue, anchoring biases, and inconsistent criterion weighting (Conway et al., 1995; Campion et al., 1997; Landy and Farr, 1980), eroding the consistency that high-stakes decisions depend on. An automated assistant addresses this directly: it applies one rubric uniformly across a large candidate pool and makes the evidence behind each score inspectable.

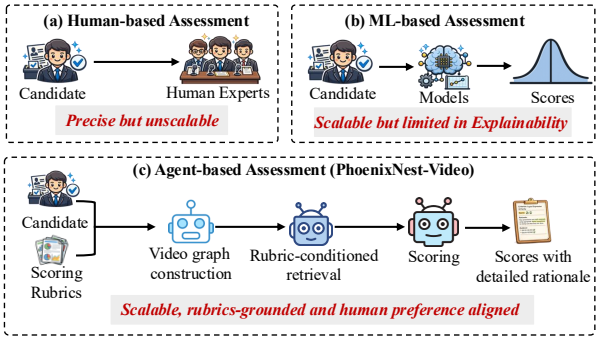


Figure 1: Three paradigms of interview assessment. (a) Human-based assessment is precise but unscalable. (b) ML-based assessment is scalable but opaque. (c) PhoenixNest-Video produces per-criterion scores with verifiable evidence, achieving scalable and evidence-traceable assessment.

As illustrated in Figure 1, interview assessment has evolved through three paradigms. Human panels remain the gold standard in accuracy but cannot scale to current applicant volumes. ML-based approaches (Naim et al., 2015; Subramaniam et al.,

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 16045–16067
October 24-29, 2026 ©2026 Association for Computational Linguistics

2016; Hemamou et al., 2019; Agrawal et al., 2020) train supervised models on interview data and gain scalability, yet they yield opaque numeric scores without interpretable rationale, limiting their utility in high-stakes decisions. Multimodal Large Language Models (MLLMs) offer a more capable foundation by processing visual, audio, and textual streams within a unified architecture (Maaz et al., 2024; Li et al., 2024b; Cheng et al., 2024; Zhang et al., 2025), yet directly applying them to interview assessment exposes structural challenges that standard training leaves unresolved.

Applying MLLMs to criterion-level evaluative judgment exposes three fundamental limitations. First, rubric scoring requires decomposing parallel visual, vocal, and verbal streams and attending to different signals per criterion (Kim et al., 2023; Arakawa and Yakura, 2022; Takeuchi and Koda, 2021), yet MLLMs treat video as an undifferentiated stream and cannot organize criterion-relevant evidence across modalities and time (Wingate and Bourdage, 2024). Second, as Figure 2 shows, general-purpose MLLMs exhibit systematic scoring biases, some compressing into a narrow low band and others skewing high; optimized for fluent generation and safety-aligned neutrality, they fail to differentiate candidates along rubric-defined dimensions. Third, defensible assessment requires every judgment to be traceable to specific behavioral observations, yet MLLMs produce assessment text without such anchoring and lack the temporal memory to cross-verify what a candidate said, showed, and expressed, leaving their outputs unauditable (Fabeyo, 2025). Fine-grained visual distinctions can also trigger hallucinations (Bai et al., 2026), while visual facts that are neither retained nor verbalized may become inaccessible in later interactions (Chen et al., 2026).

To address these challenges, we propose PhoenixNest-Video, an evidence-grounded multimodal agent framework for automated video interview assessment. PhoenixNest-Video constructs a semantic video graph as structured working memory, performs rubric-conditioned retrieval with cross-modal verification to locate and validate criterion-relevant evidence, and produces per-criterion scores anchored to the candidate's materials. A Rubrics-based Reinforcement Learning procedure with dual reward signals for rubric alignment and score-level differentiation trains the scoring module to internalize the discriminative structure of multi-level rubrics. Built on a Qwen3-VL-8B backbone, PhoenixNest-Video achieves 91.50% grade-level accuracy on VInterview-2025 with the lowest total-score MAE and Wasserstein distance among all baselines, outperforming substantially larger proprietary models, and after retraining on RecruitView (Gupta et al., 2025) attains the best rank-correlation and concordance scores macro-averaged over its 12 regression targets.

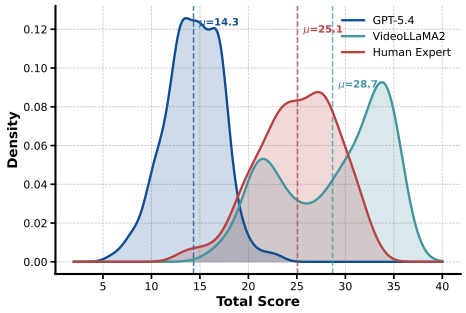


Figure 2: Score distributions of general-purpose MLLMs versus human experts on VInterview-2025. GPT-5.4 systematically underscores within a narrow band, while VideoLLaMA2 overscores with a bimodal pattern. These opposing deviations from the human expert distribution expose complementary failures in calibration and discrimination that motivate rubric-grounded reinforcement learning.

The contributions of this paper are summarized as follows:

- We propose PhoenixNest-Video, an evidence-grounded multimodal agent framework that produces criterion-level scores with verifiable video evidence, enabling transparent and auditable automated interview assessment.
- We introduce Rubrics-based Reinforcement Learning with dual reward signals for rubric alignment and score-level differentiation, enabling MLLMs to overcome scoring biases and produce rubric-faithful interview assessments.
- We evaluate PhoenixNest-Video on VInterview-2025 and RecruitView, where it achieves 91.50% grade-level accuracy and produces score distributions closely aligned with human experts, demonstrating the viability of compact, evidence-grounded multimodal agents for automated interview assessment.

2 Related Work

Automated Video Interview Assessment. Automated interview assessment has progressed from

handcrafted prosodic and facial features on mock interviews (Naim et al., 2015; Chen et al., 2016, 2017) through bimodal personality prediction on short clips (Subramaniam et al., 2016; Escalante et al., 2018) to hierarchical neural models for asynchronous screening (Hemamou et al., 2019; Agrawal et al., 2020; Singhania et al., 2020). Recent work has broadened to audio-visual personality benchmarks (Liao et al., 2024), pose-based analysis (Tang et al., 2025), multimodal performance assessment (Li et al., 2025a; Inam et al., 2026), naturalistic interview datasets (Gupta et al., 2025), fairness audits (Mujtaba and Mahapatra, 2025; Leong et al., 2019; Putra et al., 2024), psychometric validation (Liff et al., 2024), and multi-agent evaluation frameworks (Sun et al., 2026). These efforts predominantly output holistic or trait-level scores without rubric grounding or evidence traceability. PhoenixNest-Video addresses this gap by conditioning every score on rubric descriptors and anchoring it to verified candidates' materials.

MLLMs for Video Understanding. Multimodal Large Language Models (MLLMs) typically bridge a vision encoder with an LLM via instruction tuning (Maaz et al., 2024; Zhang et al., 2023), with subsequent work improving temporal modeling (Cheng et al., 2024; Zhang et al., 2025; Shao et al., 2025) and long-form processing through time-aware querying (Ren et al., 2024), dual-rate visual streams (Xu et al., 2024), and chunk-level compression (Shu et al., 2025). Reasoning-oriented post-training via reinforcement learning has further improved spatio-temporal reasoning (Feng et al., 2025; Li et al., 2025b; Tao et al., 2025). Outside video, domain-specific MLLMs have used multidimensional reasoning rewards and iterative visual tools to improve structured reasoning (Hao et al., 2026b; Fan et al., 2026a), while multimodal agents combine tool use with knowledge-grounded retrieval (Hao et al., 2026a). Recent benchmarks have also begun to test what MLLMs miss beyond surface recognition, including intent-level audiovisual understanding (Fan et al., 2026b), cross-modal ambiguity resolution (Wang et al., 2025b), and competence in humanities and social-science domains where judgment rather than fact retrieval is at stake (Kang et al., 2026b). Multi-agent orchestration provides another route to structured reasoning, both over video (Kugo et al., 2025) and in high-stakes decision domains such as legal judgment prediction (Kang et al., 2026a), though benchmarks such as Neptune (Nagrani et al., 2024) show that long-horizon temporal reasoning remains a bottleneck. In contrast to these general-purpose systems, PhoenixNest-Video targets evidence-grounded assessment and traceable reasoning over structured interview evaluations.

3 Task Formulation

Video interview assessment is widely adopted across educational, professional, and organizational contexts to evaluate candidates through structured multimodal interactions. In a typical protocol, a candidate delivers a presentation before a panel, sometimes followed by a question-and-answer session. Each panelist then independently rates the candidate across a predefined set of criteria and submits a separate score vector, and a final outcome is derived by aggregating the individual scores.

Our objective is to develop an evaluation system that emulates this multi-criterion assessment process. We define the evaluation function $\mathcal{F}$ that maps a candidate's video V, optional supplementary materials M, a set of N criteria $C = \{c_1, \ldots, c_N\}$, and their rubrics $R(c, s)$ providing a textual descriptor for every score level s in an ordinal scale, to per-criterion scores $\mathcal{S} = \{s_1, \ldots, s_N\}$, evidence references $E = \{(m_i, \text{rationale}_i)\}$ that link each judgment to specific elements of the candidate's materials, and textual feedback F

$$\mathcal{S}, E, F \leftarrow \mathcal{F}(V, M, C, R). \tag{1}$$

This formulation is general. The number of criteria N, the score scale, and the rubric content R are parameters supplied by the application setting. We evaluate on two benchmarks with different configurations, as described in Section 5.

4 Methodology

Figure 3 shows the overall architecture. Given a raw interview video and a rubric, PhoenixNest-Video first runs multimodal preprocessing (Section 4.1) to obtain structured visual and audio streams and to expand rubric descriptors into fine-grained behavioral indicators. Four modular components (Section 4.2) then produce the assessment via the Scorer policy

$$(\hat{s}, r) \sim \pi_\theta(\cdot \mid V, c, R(c, \cdot), \mathcal{B}(c, \cdot), A), \tag{2}$$

where $\hat{s}$ is the predicted score, r its rationale, and A the verified evidence chain assembled by the pipeline (Section 4.2). Only π_θ^* is trained offline by supervised fine-tuning followed by rubrics-based reinforcement learning (Section 4.3).

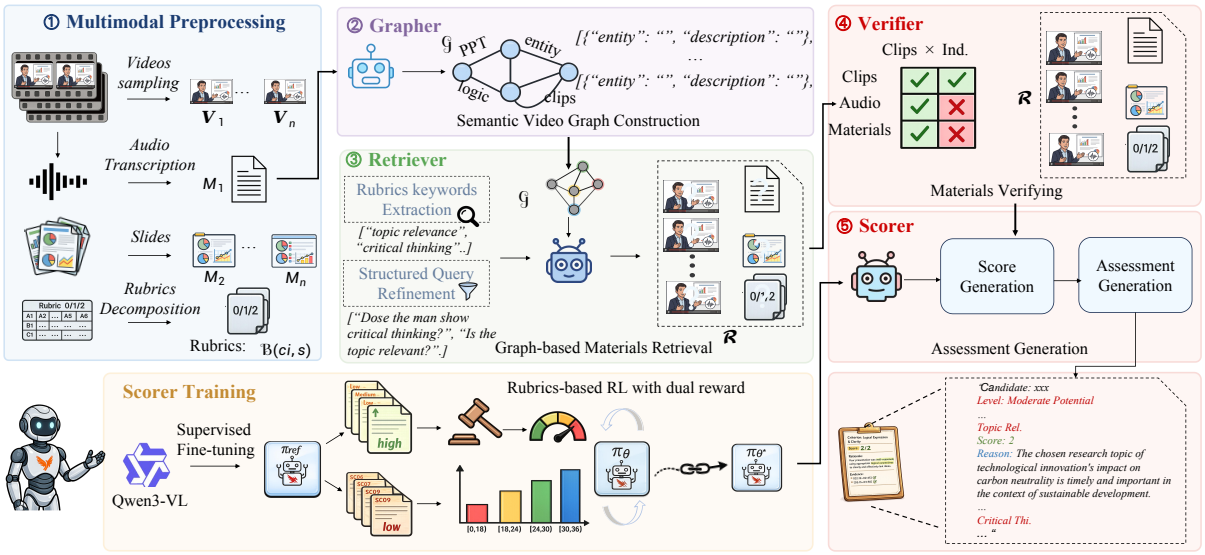


Figure 3: Overview of the PhoenixNest-Video framework. Multimodal Preprocessing converts the interview into aligned visual, audio/transcript, and slide streams, while Rubric Decomposition expands each score descriptor into behavioral indicators $\mathcal{B}(c, s)$. The Grapher constructs a clip-level semantic index; the Retriever selects criterion-relevant candidate clips; the Verifier checks the candidates across modalities; and the trained Scorer π_θ^* produces criterion-level scores, evidence references, and feedback. The Scorer is optimized by supervised fine-tuning followed by rubrics-based reinforcement learning with alignment (R_{align}) and differentiation (R_{diff}) rewards.

4.1 Multimodal Preprocessing

Before entering the agent framework, PhoenixNest-Video transforms raw video interviews into structured multimodal inputs through three parallel processing streams.

Visual Stream. We uniformly sample 32 frames per video to capture the candidate's expressions, body language, and overall presentation demeanor, with 32 chosen empirically as the accuracy peak (Section 5.5). We additionally extract slide images to capture presentation content. This global sample is criterion-agnostic and runs in parallel with the clip partitioning of the Grapher (Section 4.2). Key moments arise downstream, per criterion, as the clips that survive rubric-conditioned retrieval and multimodal verification.

Audio Stream. We extract the audio track from each video, apply ZipEnhancer for noise reduction and speech clarity enhancement, and transcribe the enhanced audio using Whisper-Large-v3 (Radford et al., 2023).

Rubric Decomposition. For each criterion $c \in C$ with rubric descriptors $R(c, s)$ for $s \in \{0, 1, 2\}$, we use an MLLM to expand the concise rubric text into fine-grained behavioral indicators:

$$\mathcal{B}(c, s) = \{b_1, b_2, \ldots, b_m\} \leftarrow \text{MLLM}(R(c, s)), \tag{3}$$

where m is the number of behavioral indicators generated for the pair (c, s), ranging from three to six in our implementation, and each b_i describes a concrete behavioral manifestation corresponding to score level s. For instance, for Critical Thinking at score level $s = 2$, the decomposition produces indicators such as "analyzes the problem from multiple complementary angles" and "supports claims with specific evidence"; at $s = 0$, it produces "provides circular or unsubstantiated reasoning" and "fails to engage with the substance of the question." These behavioral indicators ground the reward signals during training and direct the retrieval queries during inference.

4.2 Pipeline Components

To ensure that every score decision is anchored in verifiable video content, PhoenixNest-Video passes each video through four sequential stages at inference time, transforming a long-form video interview into a structured, auditable assessment through the Grapher, Retriever, Verifier, and Scorer.

Grapher. We partition the video V at 1.0 FPS into clips $\{V_1, \ldots, V_n\}$ of $K = 64$ frames each and prompt an MLLM to extract open-vocabulary semantic mentions E_i from each clip together with its aligned transcript C_i. A mention is an entity (a person, object, or presentation material, stored as a short name with a description), an action, or a scene label marking the interview phase. The vocabulary

is open and drawn from the clip content itself. The extraction prompt and output schema are given in Appendix H. Two mentions are compared by cosine similarity on L2-normalized `[CLS]` embeddings from `BAAI/bge-large-en-v1.5`,

$$\mathrm{sim}(t_a, t_b) = \frac{\mathbf{e}_a \cdot \mathbf{e}_b}{\|\mathbf{e}_a\| \, \|\mathbf{e}_b\|}, \tag{4}$$

and are merged into one prototype entity when $\mathrm{sim} > \tau$. The semantic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ has clip nodes $\mathcal{V} = \{v_i\}_{i=1}^{n}$, and an edge joins two clips whenever they share a prototype entity. Edges therefore link temporally separated moments that carry semantically equivalent content, which is what a rubric criterion typically requires: a candidate may demonstrate critical thinking in the self-introduction, again in the project presentation, and again under questioning. $\mathcal{G}$ is built once and reused across all criteria.

Retriever. For each criterion c_i, the Retriever (i) takes the behavioral indicators $\mathcal{B}(c_i, s)$ from Eq. 3 and extracts semantic keywords $\mathcal{K}$, (ii) prompts an MLLM to refine each indicator b_j into targeted queries $Q(b_j)$, and (iii) matches them against $\mathcal{G}$ under Eq. 4. A query keyword that matches a prototype entity above θ pulls in *every* clip node sharing that prototype, so the merged-entity edges act as an inverted index that surfaces temporally distant evidence a clip-local match would miss. A second pass adds nodes whose own attributes exceed θ directly, and the union is re-ranked by average embedding similarity, keeping the top-N_r candidate clips $\mathcal{R}$. $\mathcal{R}$ is topically relevant but may include false positives.

Verifier. For each indicator b_j and clip $v_i \in \mathcal{R}$, an MLLM is queried with the binary question "Does this clip demonstrate b_j?" across visual, audio, and textual modalities with equal weight; v_i is retained in $\mathcal{R}'$ if at least one modality answers positively above confidence δ. The retained clips form the reasoning chain A, a criterion-specific evidence package in which each entry records the clip identifier, its temporal range in the interview, the matched indicator b_j, and per-modality flags marking which channels supported the match. Clips on which no modality passes δ are dropped, and the surviving entries are ordered chronologically, so A reads as a time-ordered evidence trail that a reviewer can replay against the recording.

Scorer. The trained Scorer π_θ^* is invoked on the reasoning chain A, the rubric $R(c_i, \cdot)$, and the behavioral indicators $\mathcal{B}(c_i, \cdot)$ for all score levels, producing the predicted score $\hat{s}$, its rationale r, textual feedback F, and references to the candidate's materials drawn from $\mathcal{R}'$. Because π_θ^* has been optimized (Section 4.3) for rubric-faithful rationales and institutionally calibrated scores, its output is directly consumable as the final report.

4.3 Scorer Training

The Scorer component π_θ^* invoked at the assessment stage is trained offline in two steps: supervised fine-tuning followed by rubrics-based reinforcement learning.

Supervised Fine-Tuning. We fine-tune the base model on expert annotation pairs $\mathcal{D}_{\text{SFT}} = \{(V_j, c_j, s_j^*, r_j^*)\}$, establishing the basic evaluation format, output structure, and initial score distribution.

Rubrics-based Reinforcement Learning. Recent clinical MLLM work shows that continuous rubric-based rewards can provide informative supervision when binary feedback is sparse (Fan et al., 2026a), and structured reward shaping has been used more broadly to densify the learning signal in RL post-training (Shi et al., 2026). A binary exact-match GRPO reward is likewise too sparse here: with 18 criteria on a three-level scale and a total spanning $[0, 36]$, exact matches are rare in early training and most rollouts receive zero gradient. We therefore design two complementary reward signals that decompose assessment quality along orthogonal axes.

Alignment Reward. The alignment reward R_{align} measures whether the generated rationale correctly references and applies the rubric criteria for the predicted score level. We employ an independent LLM as a judge that receives the rubric descriptor $R(c, s)$, the behavioral indicators $\mathcal{B}(c, s)$, the model's predicted score $\hat{s}$, and the generated rationale r, and evaluates:

$$R_{\text{align}} = \text{LLM}_{\text{judge}}(R(c, s),\ \mathcal{B}(c, s),\ \hat{s},\ r), \tag{5}$$

where $\text{LLM}_{\text{judge}}$ assesses whether the rationale faithfully grounds its reasoning in the rubric descriptors and whether the cited behavioral evidence supports the assigned score level.

Differentiation Reward. The differentiation reward R_{diff} measures the agreement between the model's predicted total score and the expert-assigned total score at the institutional grading level. Let $L(\cdot)$ denote the level mapping function that assigns a total score to one of four institutional levels: $[0, 18)$, $[18, 24)$, $[24, 30)$, $[30, 36]$. The re-

Category	Model	Params	Total Score			Per-Criterion	
			ACC ↑	MAE ↓	W Dist ↓	QWK ↑	MAE ↓
Proprietary	GPT-5.4 (Singh et al., 2025)	-	0.4300	10.7350	10.7283	0.0618	0.6350
	Claude-opus-4-6 (Anthropic, 2025)	-	0.4450	12.9750	12.9583	0.0915	0.7456
	Gemini-3.1-pro-preview (Team et al., 2023)	-	0.8500	5.4633	4.2333	**0.1733**	0.5092
	Grok-4.1 (xAI, 2025)	-	0.9000	4.3517	2.4733	0.1069	0.4797
Open-Source	Qwen3.5-397B-A17B (Bai et al., 2025)	397B	0.7800	6.8767	6.2833	0.1172	0.4947
	Qwen3.5-27B (Bai et al., 2025)	27B	0.7800	6.9583	6.4883	0.1287	0.5267
	Kimi-K2.5 (Team et al., 2026)	1000B	0.8050	6.9217	6.4783	0.1111	0.5406
	GLM-4.5v (Hong et al., 2025)	106B	0.7688	6.9899	3.5528	0.0738	0.5137
Video-Specific	Video-R1 (Feng et al., 2025)	7B	0.7310	6.6447	3.5854	0.0223	0.4929
	VideoChat-R1 (Li et al., 2025b)	7B	0.8000	5.9983	2.7133	0.0701	0.4333
	VideoRFT (Wang et al., 2025a)	7B	0.6300	10.4867	9.1483	-0.0123	0.5850
	VideoLLaMA2 (Cheng et al., 2024)	7B	0.6782	7.0230	4.2337	0.0032	0.5160
	VideoLLaMA3 (Zhang et al., 2025)	7B	0.8528	5.6210	3.6870	-0.0179	0.4123
	PhoenixNest-Video (trained Scorer)	8B	**0.9150**	**4.1316**	**2.0328**	0.1352	0.4419
	PhoenixNest-Video (Qwen3-397B)	397B	0.8700 (Δ+0.09)	4.8450 (Δ-2.03)	3.2850 (Δ-3.00)	0.1552 (Δ+0.04)	0.4175 (Δ-0.08)
	PhoenixNest-Video (Kimi-K2.5)	1000B	0.8400 (Δ+0.04)	5.1250 (Δ-1.80)	3.8650 (Δ-2.61)	0.1689 (Δ+0.06)	0.4219 (Δ-0.12)
	PhoenixNest-Video (GLM-4.5v)	106B	0.8950 (Δ+0.13)	4.4150 (Δ-2.57)	2.6350 (Δ-0.92)	0.1171 (Δ+0.04)	**0.3869** (Δ-0.13)

Table 1: Main results on the VInterview-2025 test set. **Bold** indicates the best result per column; underline indicates the second best. "Proprietary" refers to closed-source general-purpose models accessed via commercial APIs; "Open-Source" refers to publicly available general-purpose open models; "Video-Specific" refers to models specifically designed or optimized for video understanding and reasoning. Total score metrics measure alignment at the institutional grading level; per-criterion metrics measure agreement on individual 0–2 rubric scores. Missing or non-parseable model outputs are excluded from scoring, so the number of scored interviews is at most 200 and varies across models. For the three backbone-amplifier rows in the *Ours* block, the small green value in parentheses, e.g. (Δ+0.09), denotes the absolute change of the wrapped configuration relative to the same backbone's direct-prompt counterpart in the *Open-Source* block; the sign already encodes the direction of improvement with respect to the metric.

ward is:

$$R_{\text{diff}} = \begin{cases} 1.0 & \text{if } L(\hat{S}_{total}) = L(S^*_{total}), \\ 0.5 & \text{if } |L(\hat{S}_{total}) - L(S^*_{total})| = 1, \\ 0.0 & \text{if } |L(\hat{S}_{total}) - L(S^*_{total})| \geq 2, \end{cases} \tag{6}$$

where i indexes the rubric criteria, $\hat{S}_{total} = \sum_{i=1}^{N} \hat{s}_i$ is the model's predicted total score summed over the N criterion-level predictions $\hat{s}_i$, and S^*_{total} is the expert total score.

The total reward is $R = \lambda_1 R_{\text{align}} + \lambda_2 R_{\text{diff}}$, and we optimize π_θ with standard GRPO (Shao et al., 2024) using group-normalized advantages and a KL penalty $\beta\,\mathrm{KL}(\pi_\theta \| \pi_{\text{ref}})$ against the reference policy.

5 Experiments

We design experiments around four research questions that together evaluate whether PhoenixNest-Video meets the practical demands of an admissions-facing copilot:

- **RQ1** How does PhoenixNest-Video compare with proprietary, open-source, and video-specialized baselines on real-world video interview assessment, and does the framework lift existing MLLM backbones beyond their direct-prompt performance?
- **RQ2** What is the individual contribution of each key component, namely the alignment reward R_{align}, the differentiation reward R_{diff}, and the retrieval-and-verification pipeline formed by the Grapher, Retriever, and Verifier?
- **RQ3** Does the framework remain effective when retrained under a different annotation schema, i.e., on a benchmark whose targets are continuous personality and performance dimensions rather than rubric-grounded ordinal scores?
- **RQ4** Does PhoenixNest-Video exhibit any systematic bias along sensitive demographic or disciplinary axes, and how sensitive are its scores to the number of sampled frames per video?

5.1 Experimental Setup

Datasets. We evaluate on two complementary benchmarks. **VInterview-2025** is a self-collected dataset of 491 real-world graduate admissions interviews recorded during a live admissions cycle at the participating institution. The recordings were obtained passively from the existing admissions workflow and at no point fed back into the decision process, so the panel scores reflect a genuine high-stakes evaluation rather than an annotation task performed for our work; this yields high-quality rubric labels while keeping the data collection free

Model	Spearman ρ	Kendall τ-b	C-index	Pearson r
GPT-5.4 (Singh et al., 2025)	0.3096	0.2218	0.6067	0.2702
Claude Opus 4.6 (Anthropic, 2025)	<u>0.3832</u>	<u>0.2763</u>	<u>0.6328</u>	<u>0.3756</u>
Gemini-3.1-Pro-Preview (Team et al., 2023)	0.1775	0.1303	0.5571	0.1560
Grok-4.1 (xAI, 2025)	0.2962	0.2075	0.6022	0.2437
Doubao-Seed-1.8-251228 (Guo et al., 2025)	0.2985	0.2158	0.5979	0.2440
Qwen3.5-397B-A17B (Bai et al., 2025)	0.2973	0.2120	0.6018	0.2559
GLM-4.5v (Hong et al., 2025)	0.3313	0.2350	0.6141	0.2926
Kimi-K2.5 (Team et al., 2026)	0.1914	0.1436	0.5591	0.1849
PhoenixNest-Video	**0.4437**	**0.3159**	**0.6579**	**0.4234**

Table 2: Macro-averaged performance on the RecruitView dataset (Gupta et al., 2025) across all 12 targets on the 317-sample user-grouped test split. **Bold** indicates the best result and <u>underline</u> the second best in each column. All baseline rows report our own evaluation of the corresponding model under the prompt-based protocol. Missing or non-parseable API outputs are excluded from scoring.

of any decision-altering intervention. Each interview lasts approximately 15 minutes and is rated by a three-member faculty panel on 18 rubric criteria (0–2 scale). The rubric is the institution's standard admissions instrument, maintained by the admissions committee and in operational use across earlier cycles; it was not written or modified for this study. All three panelists are faculty with admissions experience who received institutional training on the criteria and scoring standards, and each submits a separate 18-criterion score vector. Following the institution's own aggregation rule, the total-score ground truth is the arithmetic mean of the three faculty totals, $S_{\mathrm{GT}} = \frac{1}{3}\sum_{r=1}^{3} S^{(r)}$, and the per-criterion reference is the arithmetic mean of the three faculty criterion scores. We use 100 interviews for supervised fine-tuning, 191 for rubrics-based reinforcement learning, and hold out 200 for testing; the two training stages draw on non-overlapping candidates, and the 200 test interviews come from a different admissions round than the training data, so the evaluation is cross-round rather than a random split within one batch (Appendix B.2). To protect candidate privacy, the corpus stays on institutional infrastructure and is not publicly released. **RecruitView** (Gupta et al., 2025) is a public benchmark of 2,011 job-interview videos with expert annotations along 12 regression targets covering overall personality, speaking skills, confidence, the Big Five personality traits, and additional interview-performance dimensions, with the official user-stratified split of 1,404 training, 290 validation, and 317 test clips, on which we retrain and evaluate our framework. **Baselines.** We compare against three categories of MLLMs under the same prompt-based protocol: four proprietary (Singh et al., 2025; Anthropic, 2025; Team et al., 2023; xAI, 2025), four open-source (Bai et al., 2025; Team et al., 2026; Hong et al., 2025), and five video-specialized (Feng et al., 2025; Li et al., 2025b; Wang et al., 2025a; Cheng et al., 2024; Zhang et al., 2025).

Implementation Details. We use Qwen3-VL-8B-Instruct as the backbone, trained for 3 epochs with a batch size of 16 and a learning rate of 2×10^{-5}. The LLM judge for R_{align} is GPT-5.4. Reward weights are $\lambda_1 = \lambda_2 = 0.5$ with $\beta = 0.05$. The retrieval-and-verification pipeline uses the untrained base model for graph construction and retrieval, while the final assessment is generated by the trained π_θ. We use VLMEvalKit (Duan et al., 2024) to execute model inference and standardize output parsing during evaluation. Video processing uniformly samples 32 frames per video, while the Grapher partitions the video at 1.0 FPS into $K = 64$-frame clips; entity matching uses `BAAI/bge-large-en-v1.5` embeddings with merging threshold $\tau = 0.7$ and retrieval threshold $\theta = 0.5$. Appendix C.2 reports the ablations over the embedding model and θ that motivate this configuration. All experiments run on $4\times$A800 GPUs (80GB).

Metrics. Following prior work (Ke et al., 2024; Gu et al., 2024; Li et al., 2024a; Gupta et al., 2025), we report grade-level accuracy (ACC, computed over the institution's four-tier grading scheme, which aggregates raw rubric scores into the admissions-

relevant grades faculty reviewers act on; a prediction counts as correct when its tier is within one tier of the expert tier, as defined in Appendix D), mean absolute error (MAE), Wasserstein distance (W Dist) at the total-score level, and Quadratic Weighted Kappa (QWK) at the per-criterion level for VInterview-2025 test set. For RecruitView (Gupta et al., 2025), we report Spearman ρ, Kendall τ-b, C-index, and Pearson r, macro-averaged over all 12 targets following the evaluation protocol of the original paper.

5.2 Main Results (RQ1)

Comparison with baselines. Table 1 reports results on the VInterview-2025 test set. PhoenixNest-Video achieves the best total-score alignment and remains competitive on per-criterion agreement, despite using only an 8B backbone. Among the baselines, the two strongest proprietary models, Gemini-3.1-pro-preview and Grok-4.1, lead every other baseline on grade-level accuracy, while GPT-5.4 and Claude-opus-4-6 fall below the open-source and video-specialized groups; video-specialized models tend to be stronger on fine-grained per-criterion MAE. The fact that our 8B configuration matches or surpasses substantially larger proprietary and open-source models indicates that rubric-grounded training and evidence-grounded retrieval are more effective than raw scale for structured assessment.

Framework as a backbone amplifier. The remaining *Ours* rows wrap three large MLLMs (Qwen3.5-397B-A17B, Kimi-K2.5, GLM-4.5v) inside the PhoenixNest-Video framework. Each configuration consistently improves over the same backbone's direct-prompt counterpart on every metric, as shown by the green deltas next to each value. The largest absolute gain appears for GLM-4.5v, whose grade-level accuracy increases by more than ten points and whose per-criterion MAE drops by a comparable margin; Qwen3.5-397B-A17B and Kimi-K2.5 show similar trends. This confirms that the improvements stem from the framework itself rather than from a particular choice of backbone, supporting the claim that PhoenixNest-Video acts as a backbone-agnostic capability amplifier for existing MLLMs.

5.3 Ablation Studies (RQ2)

Figure 4 isolates the contribution of each component by removing one at a time from the full framework. Removing either reward signal causes

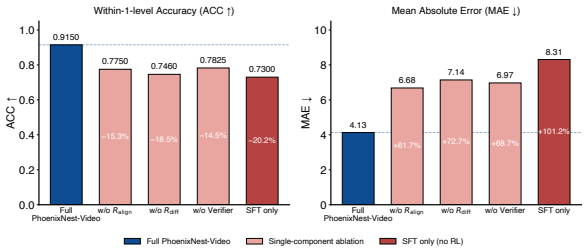


Figure 4: Ablation of PhoenixNest-Video on VInterview-2025. Removing any single component or skipping reinforcement learning consistently degrades grade-level accuracy and inflates MAE, indicating that the rubrics-based rewards and the retrieval-and-verification pipeline are complementary.

a substantial drop in accuracy and a sharp rise in MAE, with the differentiation reward R_{diff} being slightly more impactful than the alignment reward R_{align}. This pattern indicates that the two rewards are complementary along orthogonal axes: R_{align} enforces rubric-faithful rationales while R_{diff} shapes the harder institutional-level score differentiation. Removing the Verifier, so that every retrieved clip is passed to the Scorer without cross-modal checking, produces a comparable degradation, confirming that verification contributes on top of the reward shaping rather than being subsumed by it. Finally, the SFT-only baseline yields the largest overall regression, validating that supervised fine-tuning alone is insufficient and that reinforcement training with structured rewards is necessary to internalize the rubric.

5.4 Results on RecruitView (RQ3)

To verify that the framework is not tied to the self-collected rubric schema, we retrain and evaluate PhoenixNest-Video on the RecruitView dataset (Gupta et al., 2025) following its native train/test protocol. RecruitView differs from VInterview-2025 in two fundamental ways, its targets are continuous regression scores rather than 0–2 ordinal rubric scores, and they cover the Big Five personality traits, an overall-personality index, and six interview-performance dimensions rather than presentation- and Q&A-oriented criteria. We compare against the proprietary and open-source models of Table 1, plus Doubao-Seed-1.8 (Guo et al., 2025), under the same prompt-based protocol. As Table 2 shows, PhoenixNest-Video attains the best score across all four rank-correlation and concordance metrics, while every baseline remains at moderate correlation levels, underscoring the intrin-

Gender	Mean Total Score
Male	25.86
Female	26.52
Δ (Female $-$ Male)	$+0.66$ (1.8%)

Table 3: Mean predicted total score (max 36) by candidate gender on VInterview-2025. The aggregate gap is small relative to the score range and within-group variance.

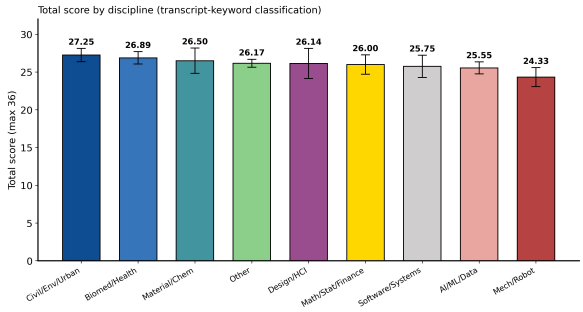


Figure 5: Mean predicted total score (max 36) by candidate discipline on VInterview-2025, where each candidate is assigned directly to their reported academic discipline. Error bars denote 95% bootstrap confidence intervals.

sic difficulty of video interview assessment even for strong proprietary models. The consistent advantage indicates that the rubric-grounded training procedure and the retrieval-and-verification pipeline are not coupled to the specific labeling convention on which they were originally developed, and that they remain effective when the target structure changes.

5.5 Deep Analysis (RQ4)

Beyond aggregate accuracy, we further analyze two properties of PhoenixNest-Video on VInterview-2025. First, we audit predictions for systematic score gaps along two sensitive axes, candidate gender and academic discipline. Second, we examine the sensitivity of the framework to the number of sampled frames per video, which is the most consequential preprocessing hyperparameter in our pipeline.

Gender Bias. Table 3 reports the mean predicted total score on VInterview-2025 by candidate gender. The gap is small relative to the score range and the within-group spread, and the scores show no systematic preference for either gender. The Limitations section defines the scope of this audit.

Subject Bias. Figure 5 groups candidates by their academic discipline. The per-discipline means span a moderate range, with engineering-heavy disciplines slightly below the cross-discipline mean and verbal- or scenario-heavy disciplines slightly above. We attribute the residual spread to differences in evidence density: applicants in disciplines with richer verbal content tend to surface more rubric-aligned behavioral indicators during retrieval and verification. The overlap of per-discipline confidence intervals indicates that PhoenixNest-Video spreads its scores across fields rather than favoring a few.

Sampled Frames	ACC
10	0.870
16	0.895
32	**0.915**
64	0.885

Table 4: Effect of the number of uniformly sampled frames per video on grade-level accuracy on VInterview-2025. Accuracy peaks at 32 frames and then declines, indicating that denser sampling is not always beneficial.

Effect of Frame Sampling Density. Table 4 reports grade-level accuracy as we vary the number of uniformly sampled frames per video from 10 to 64. Accuracy rises sharply from 10 to 16 frames, plateaus through 32, and degrades at 64. Too few frames omit criterion-relevant evidence, whereas too many inflate the visual token budget and dilute attention across redundant near-duplicate frames. We therefore adopt 32 frames as the default throughout this paper.

6 Conclusion

We presented PhoenixNest-Video, a multimodal agent framework for automated video interview assessment. Rubrics-based Reinforcement Learning supplies dual rewards for rubric alignment and score-level differentiation, and a four-stage agent pipeline anchors each score to the elements of the candidate's materials that survive cross-modal verification. On VInterview-2025 and RecruitView, PhoenixNest-Video outperforms substantially larger proprietary and open-source baselines, showing that rubric grounding and evidence traceability matter more than raw scale for structured assessment.

Limitations

Four conditions define where these results apply.

Scope of the fairness audit. The audit covers can-

didate gender and academic discipline in a single admissions cohort at one institution. Within these groups and this sample, predicted scores show no large aggregate difference, and the per-group sizes make only large disparities detectable. The admissions workflow the data comes from records no other candidate attributes, so ethnicity, nationality, accent, disability, socioeconomic background, and first-language background fall outside the audit; collecting them would require consent and data-protection provisions beyond the present ethical-review approval. The behavioral and linguistic signals the framework retrieves may also track culturally specific communication norms, or unequal access to interview coaching and recording conditions, rather than candidate capability.

Language proficiency and transcript quality. The interviews are conducted in English and many candidates are non-native speakers, which constrains the system in two ways. First, ASR accuracy degrades under strong accents, background noise, and low-quality recordings, and these errors propagate into the transcript stream the Grapher and the Retriever consume. Measuring this effect requires human reference transcripts, which VInterview-2025 does not include. Second, disfluent or error-prone speech can depress criteria that target content rather than language, so a candidate with limited fluency may lose points for a reason the rubric does not intend to measure. The Scorer is trained on faculty ratings and reproduces this confound. Separating language proficiency from the competencies the rubric targets calls for interview data with wider variation in recording conditions and speaker language background.

Aggregated ground truth. The reference score is the equal-weight mean of three faculty score vectors, the rule the institution applies in its own admissions decisions. This aggregation preserves panel disagreement. For candidates on whom the faculty diverge, the reference is a value that no individual reviewer assigned, so agreement with the reference measures agreement with the panel mean rather than with any single expert.

Fixed design choices and deployment scope. Two components are set once and not ablated. The entries of the reasoning chain A reach the Scorer in chronological order, and LLMs are sensitive to the ordering of structured inputs (He et al., 2026), so an alternative ordering such as by verification confidence may shift the resulting scores. The pipeline also assumes that the full interview is available before assessment begins; extending it to streaming video, so that intermediate scores and evidence anchors appear while the interview unfolds, remains open. VInterview-2025 itself stays closed. The recordings are graduate admissions interviews with identifiable candidates, and no anonymization we can apply to video and voice removes that identifiability, so we keep the corpus on institutional infrastructure and release neither the recordings nor the derived benchmark.

References

Anumeha Agrawal, Rosa Anil George, Selvan Sunitha Ravi, Sowmya Kamath S, and Anand Kumar. 2020. Leveraging multimodal behavioral analytics for automated job interview performance assessment and feedback. *ArXiv*, abs/2006.07909.

Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A comprehensive review of ai techniques for addressing algorithmic bias in job hiring. *Ai*, 5(1):383–404.

Philip G Altbach, Liz Reisberg, and Laura E Rumbley. 2019. *Trends in global higher education: Tracking an academic revolution*, volume 22. Brill.

Anthropic. 2025. Introducing claude sonnet 4.5. *https://www.anthropic.com/news/claude-sonnet-4-5*.

Riku Arakawa and Hiromu Yakura. 2022. Ai for human assessment: What do professional assessors need? *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Tianyi Bai, Yuxuan Fan, Qiu Jiantao, Fupeng Sun, Jiayi Song, Junlin Han, Zichen Liu, Conghui He, Wentao Zhang, and Binhang Yuan. 2026. Hallucination at a glance: Controlled visual edits and fine-grained multimodal learning. *Advances in Neural Information Processing Systems*, 38:135360–135393.

Michael A Campion, David K Palmer, and James E Campion. 1997. A review of structure in the selection interview. *Personnel psychology*, 50(3):655–702.

Hong Chen, Kang Chen, Yuxuan Fan, Bo Wang, Yubo Gao, Yuanlin Chu, and Xuming Hu. 2026. Seen, said, or forgotten? a causal audit of visual kv memory across dialog turns. *arXiv preprint arXiv:2607.25467*.

Lei Chen, Gary Feng, Michelle P Martin-Raugh, Chee Wee Leong, Christopher Kitchen, Su-Youn Yoon, Blair Lehman, Harrison Kell, and Chong Min

Lee. 2016. Automatic scoring of monologue video interviews using multimodal cues. In *INTERSPEECH*, pages 32–36.

Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 504–509. IEEE.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

James M Conway, Robert A Jako, and Deborah F Goodman. 1995. A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of applied psychology*, 80(5):565.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11198–11201.

HJ Escalante, H Kaya, AA Salah, S Escalera, Y Gucluturk, U Guclu, X Baró, I Guyon, JJ Junior, M Madadi, et al. 2018. Explaining first impressions: modeling, recognizing, and explaining apparent personality from videos. arxiv preprint arxiv: 180200745.

Stephen Fabeyo. 2025. Explainable ai in employment decision-making: a systematic review of transparency methods in hiring algorithms. *Issues in Information Systems*, 26(3).

Yuxuan Fan, Jing Hao, Hong Chen, Jiahao Bao, Yihua Shao, Yuci Liang, Kuo Feng Hung, and Hao Tang. 2026a. Oralgpt-plus: Learning to use visual tools via reinforcement learning for panoramic x-ray analysis. *arXiv preprint arXiv:2603.06366*.

Yuxuan Fan, Gyusik Seo, Jing Hao, Jaemin Cho, Mohit Bansal, and Jaehong Yoon. 2026b. Musebench: Benchmarking intent-level audiovisual arts understanding in mllms. *arXiv preprint arXiv:2606.30026*.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. 2025. Video-r1: Reinforcing video reasoning in mllms. *arXiv preprint arXiv:2503.21776*.

Trevor Franklin. 2024. Work in progress: Development of a taxonomy of undergraduate engineering admissions practices and protocols. In *2024 ASEE Annual Conference & Exposition*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Amit Kumar Gupta, Farhan Sheth, Hammad Shaikh, Dheeraj Kumar, Angkul Puniya, Deepak Panwar, Sandeep Chaurasia, and Priya Mathur. 2025. Recruitview: A multimodal dataset for predicting personality and interview performance for human resources applications. *arXiv preprint arXiv:2512.00450*.

Jing Hao, Siyuan Dai, Yongxin Zhang, Yuci Liang, Jiamin Wu, Jiahao Bao, Yuxuan Fan, Zanting Ye, Yanpeng Sun, Xinyu Zhang, et al. 2026a. Oralagent: Integrating reasoning, tools, and knowledge for interactive dental image analysis. *arXiv preprint arXiv:2605.27378*.

Jing Hao, Yuci Liang, Lizhuo Lin, Yuxuan Fan, Wenkai Zhou, Kaixin Guo, Zanting Ye, Yanpeng Sun, Xinyu Zhang, Yanqi Yang, et al. 2026b. Oralgpt-omni: A versatile dental multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 38509–38519.

Yingjie He, Zhaolu Kang, Kehan Jiang, Qianyuan Zhang, Jiachen Qian, Chunlei Meng, Yujie Feng, Yuan Wang, Jiabao Dou, Aming Wu, et al. 2026. How order-sensitive are llms? orderprobe for deterministic structural reconstruction. *arXiv preprint arXiv:2601.08626*.

Léo Hemamou, G. Felhi, Vincent Vandenbussche, Jean-Claude Martin, and Chloé Clavel. 2019. Hirenet: A hierarchical attention model for the automatic analysis of asynchronous video job interviews. In *AAAI Conference on Artificial Intelligence*.

GLM-V Team Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, and 58 others. 2025. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning.

Syed Azeem Inam, Abdul Kabeer, Muneeb Ahmed Abbasi, and Abdullah Ayub Khan. 2026. Ivas: A multimodal ai system for objective video interview assessment with facial emotion, gaze, and audio analysis. *Edelweiss Applied Science and Technology*, 10(1):525–543.

Zhaolu Kang, Junhao Gong, Qingxi Chen, Hao Zhang, Jiaxin Liu, Rong Fu, Zhiyuan Feng, Yuan Wang, Simon Fong, and Kaiyue Zhou. 2026a. Multimodal multi-agent empowered legal judgment prediction.

In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12202–12206. IEEE.

Zhaolu Kang, Junhao Gong, Jiaxu Yan, Wanke Xia, Yian Wang, Zhuo Cheng, Wenhao Cao, Ziwen Wang, ZhiYuan Feng, Huaxuan Ding, et al. 2026b. Hss-bench: Benchmarking humanities and social sciences ability for multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pages 74664–74719.

Pei Ke, Bosi Wen, Andrew Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. 2024. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13034–13054.

Changwoo Kim, Jinho Choi, Jong-Soeb Yoon, Daehun Yoo, and Woojin Lee. 2023. Fairness-aware multimodal learning in automatic video interview assessment. *IEEE Access*, 11:122677–122693.

Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, et al. 2025. Videomultiagents: A multi-agent framework for video question answering. *arXiv preprint arXiv:2504.20091*.

Frank J Landy and James L Farr. 1980. Performance rating. *Psychological bulletin*, 87(1):72.

Chee Wee Leong, Katrina Roohr, Vikram Ramanarayanan, Michelle P Martin-Raugh, Harrison Kell, Rutuja Ubale, Yao Qian, Zydrune Mladineo, and Laura McCulla. 2019. To trust, or not to trust? a study of human bias in automated video interview assessments. *arXiv preprint arXiv:1911.13248*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv:2411.16594*.

Deng Li, Xin Liu, Bohao Xing, Baiqiang Xia, Yuan Zong, Bihan Wen, and Heikki Kälviäinen. 2024b. Eald-mllm: Emotion analysis in long-sequential and de-identity videos with multi-modal large language model. *arXiv preprint arXiv:2405.00574*.

Jia Li, Yang Wang, Wenhao Qian, Jialong Hu, Zhenzhen Hu, Richang Hong, and Meng Wang. 2025a. Listening to the unspoken: Exploring'365'aspects of multimodal interview performance assessment. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 13909–13916.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025b. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*.

Rongfan Liao, Siyang Song, and Hatice Gunes. 2024. An open-source benchmark of deep learning models for audio-visual apparent and self-reported personality recognition. *IEEE Transactions on Affective Computing*, 15(3):1590–1607.

Josh Liff, Nathan Mondragon, Cari Gardner, Christopher J Hartwell, and Adam Bradshaw. 2024. Psychometric properties of automated video interview competency assessments. *Journal of Applied Psychology*, 109(6):921.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602.

Jolene M Maude and Dale Kirby. 2022. Holistic admissions in higher education: a systematic literature review. *Journal of Higher Education Theory and Practice*, 22(8).

Dena F Mujtaba and Nihar R Mahapatra. 2025. Behind the screens: Uncovering bias in ai-driven video interview assessments using counterfactuals. *arXiv preprint arXiv:2505.12114*.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, et al. 2024. Neptune: The long orbit to benchmarking long video understanding. *arXiv preprint arXiv:2412.09582*.

Iftekhar Naim, Md. Iftekhar Tanveer, Daniel Gildea, and Ehsan Hoque. 2015. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9:191–204.

Bimasena Putra, Kurniawati Azizah, Candy Olivia Mawalim, Ikhlasul Akmal Hanif, Sakriani Sakti, Chee Wee Leong, and Shogo Okada. 2024. Magbert-arl for fair automated video interview assessment. *IEEE Access*, 12:145188–145205.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323.

Yihua Shao, Haojin He, Sijie Li, Siyu Chen, Xinwei Long, Fanhu Zeng, Yuxuan Fan, Muyang Zhang, Ziyang Yan, Ao Ma, et al. 2025. Eventvad: Training-free event-aware video anomaly detection. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2586–2595.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Qiming Shi, Zhaolu Kang, Yunfan Zhou, Di Weng, and Yingcai Wu. 2026. Spader: Step-wise peer advantage with diversity-aware exploration rewards for multi-answer question answering. *arXiv preprint arXiv:2606.00593*.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. 2025. Video-xl: Extra-long vision language model for hour-scale video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26160–26169.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Abhishek Singhania, Abhishek Unnam, and Varun Aggarwal. 2020. Grading video interviews with fairness considerations. *arXiv preprint arXiv:2007.05461*.

Park Song, Nolan Sarah, Kim Brian Heseung, Kajikawa Trent, Freeman Mark, Hughes Rodney, and Walter Herring. 2026. Deadline update, 2025–2026: First-year application trends through january 1.

Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. 2016. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. *ArXiv*, abs/1610.10048.

Gengxin Sun, Ruihao Yu, Liangyi Yin, Yunqi Yang, Bin Zhang, and Zhiwei Xu. 2026. Comai: A collaborative multi-agent framework for robust and equitable interview evaluation. *arXiv preprint arXiv:2603.16215*.

Nao Takeuchi and Tomoko Koda. 2021. Initial assessment of job interview training system using multimodal behavior analysis. *Proceedings of the 9th International Conference on Human-Agent Interaction*.

Bin Tang, Ke-Qi Pan, Miao Zheng, Ning Zhou, Jia-Lu Sui, Dandan Zhu, Cheng-Long Deng, and Shu-Guang Kuai. 2025. Pose as a modality: A psychology-inspired network for personality recognition with a new multimodal dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1538–1546.

Sicheng Tao, Jungang Li, Yibo Yan, Junyan Zhang, Yubo Gao, Hanqian Li, ShuHang Xun, Yuxuan Fan, Hong Chen, Jianxiang He, et al. 2025. Moss-chatv: Reinforcement learning with process reasoning reward for video temporal reasoning. *arXiv preprint arXiv:2509.21113*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, et al. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Haoxu Wang and Biao Tian. 2025. Zipenhancer: dual-path down-up sampling-based zipformer for monaural speech enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Qi Wang, Yanrui Yu, Ye Yuan, Rui Mao, and Tianfei Zhou. 2025a. Videorft: Incentivizing video reasoning capability in mllms via reinforced fine-tuning. *arXiv preprint arXiv:2505.12434*.

Xiaolong Wang, Zhaolu Kang, Wangyuxuan Zhai, Xinyue Lou, Yunghwei Lai, Ziyue Wang, Yawen Wang, Kaiyu Huang, Yile Wang, Peng Li, et al. 2025b. Mucar: Benchmarking multilingual cross-modal ambiguity resolution for multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15037–15059.

Arthur N Wiens. 1976. The assessment interview.

Timothy G. Wingate and Joshua S. Bourdage. 2024. What are interviews for? a qualitative study of employment interview goals and design. *Human Resource Management*.

xAI. 2025. Grok 4. *https://x.ai/news/grok-4*.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. 2025. Video-llama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.

A Ethical Considerations and Responsible Use

PhoenixNest-Video is a research prototype for evidence-grounded video interview assessment. Because the application domain is sensitive, we consolidate the relevant practices around data collection, anonymization, fairness, intended use, and artifact licensing in this section so that they can be reviewed together.

A.1 Anonymization and Informed Consent

All video interviews used in this work were collected in close cooperation with the participating institution and under institutional ethical-review approval. Each participant was informed in advance that their interview recording could be used for research on automated evaluation tools, and explicit written consent was obtained from every participant before any recording was retained for research use. Participation was voluntary and did not affect any admissions outcome.

Prior to model training and analysis, the recordings were processed through an anonymization pipeline that removes or replaces directly identifying information. Specifically, names, contact details, institutional affiliations, and other personally identifying entities mentioned in the transcripts were detected by a combination of automated entity recognition and manual verification, and replaced with neutral placeholders. Slides and supporting visual materials were inspected and any visible identifiers were blurred or removed. The resulting research-use copies of the recordings are stored on access-controlled institutional infrastructure. Training, graph construction, retrieval, and verification run entirely on institutional hardware. The only material that leaves this infrastructure is the anonymized frames and transcripts sent to the commercial APIs of the proprietary baselines and of the LLM judge, under the providers' terms of service; no raw, non-anonymized recording is transmitted at any stage.

A.2 Fairness Auditing

As discussed in Section 5.5, we audit PhoenixNest-Video for systematic differences in predicted scores across candidate gender and academic discipline. The observed gaps are small relative to within-group variance and to the overall score range. Gender and discipline are the only two attributes the source admissions workflow records, which fixes the coverage of the audit; the Limitations section states the consequences. Auditing is an ongoing activity: use on a new candidate population calls for a refreshed audit, with attention to subgroups that are under-represented in the historical data, and the audit results belong alongside the aggregate performance metrics.

A.3 Intended Use and Scope

PhoenixNest-Video helps faculty reviewers organize evidence and surface rubric-aligned observations from long video interviews. Admissions outcomes remain a human decision. Three conditions define the intended use: (i) the framework is built for research and benchmarking, and institutional use begins with an internal review of fitness for purpose; (ii) the output is a draft assessment that a qualified reviewer verifies against the underlying video evidence before any decision; (iii) the evidence citations produced by the Retriever and Verifier exist to make that verification fast.

A.4 License of Artifacts

We use all third-party models and datasets only for non-commercial research purposes and in accordance with their respective licenses and terms of use. The proprietary MLLMs we evaluate (GPT-5.4, Claude-Opus-4-6, Gemini-3.1-pro-preview, Grok-4.1, Doubao-Seed-1.8) are accessed through their official APIs subject to the providers' terms of service, and no proprietary model weights are redistributed by us. The open-source MLLMs and video-specialized models (Qwen3.5-397B-A17B, Qwen3.5-27B, Kimi-K2.5, GLM-4.5v, Video-R1, VideoChat-R1, VideoRFT, VideoLLaMA2, VideoLLaMA3) are used under their published licenses; we do not modify or redistribute their weights. The RecruitView benchmark (Gupta et al., 2025) is used following the access terms and evaluation protocol described in the original release. The VInterview-2025 dataset constructed in this work is not publicly released. The recordings show identifiable candidates in a high-stakes admissions setting, so they remain on access-controlled institutional infrastructure and are used only for the research reported here.

B Datasets

B.1 VInterview-2025 Description

The VInterview-2025 dataset is self-collected from real-world graduate admissions interviews at the

participating institution. Recording occurred passively during the institution's live admissions cycle, and neither the recordings nor any system output was returned to the panel; admissions decisions were made independently of this work, so the dataset captures genuine high-stakes panel judgments rather than annotations performed post hoc. This non-interventional design also keeps the data collection within the scope of the institutional ethical-review approval and avoids any influence on the candidates' outcomes. Each video interview lasts approximately 15 minutes and is organized into two phases, a presentation phase followed by a question-and-answer phase, so that both prepared exposition and live interaction are captured for evaluation. All videos are rated by a three-member faculty panel. Every panelist is a professor at the participating institution with professional expertise in candidate assessment and received institutional training on the rubric criteria and scoring standards before the admissions cycle. Each panelist independently scores all 18 rubric criteria on a three-level ordinal scale ranging from 0 to 2 and submits a separate score vector; residual disagreement between panelists is retained through equal-weight averaging rather than adjudicated away. The complete rubric used for annotation is provided in Appendix F.

B.2 VInterview-2025 Collection and Preprocessing

All VInterview-2025 recordings are collected under informed consent and institutional IRB approval, and personal identifiers are removed prior to any processing, as described in Appendix A.1. The audio stream is first enhanced by ZipEnhancer (Wang and Tian, 2025) and then transcribed by Whisper (Radford et al., 2023) to obtain time-aligned text aligned with the video timeline. All videos are normalized to a uniform spatial resolution and frame rate before being passed to the semantic video graph construction pipeline. **Stage-wise data use.** Table 5 gives the exact partition of the 491 interviews across the two training stages and the test set. The supervised fine-tuning and reinforcement learning stages draw on disjoint candidate pools, and the 200 test interviews are taken from an admissions round that contributes no training data, so the reported numbers reflect cross-round rather than within-batch generalization. We use fixed epoch budgets for both stages and perform no checkpoint selection on held-out data, so no validation split is carved out of VInterview-2025.

Stage	Interviews	Source round
SFT	100	training rounds
GRPO	191	training rounds
Test	200	separate round
Total	491	

Table 5: Stage-wise use of the 491 VInterview-2025 interviews. The SFT and GRPO pools are non-overlapping, and the test round contributes no training data.

B.3 RecruitView Description and Protocol

To test whether PhoenixNest-Video generalizes beyond the graduate-admissions setting of VInterview-2025, we additionally retrain and evaluate it on the RecruitView benchmark (Gupta et al., 2025), which targets the professional hiring scenario. RecruitView contains 2,011 in-the-wild interview clips from 300+ participants responding to 76 curated questions. Each clip carries 12 continuous, near-zero-mean target scores produced by a nuclear-norm-regularized Multinomial-Logit model fitted to about 27,000 pairwise judgments by clinical psychologists: the Big-Five personality traits (Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism) and an Overall-Personality index, together with six interview-performance dimensions (Interview Score, Answer Score, Speaking Skills, Confidence, Facial Expression, Overall Performance). We follow the official user-stratified split with 1,404 training, 290 validation, and 317 test clips, and adopt the rank-correlation and concordance metrics in Appendix D for evaluation.

Quartile-based 4-tier adapter. The original targets are continuous, whereas our reward design assumes ordinal scores. We therefore introduce a lightweight adapter that, for each dimension c, discretizes the continuous label y into a four-level ordinal score $\tilde{s} \in \{0, 1, 2, 3\}$ using the training-set quartiles $Q_{25}^{(c)}, Q_{50}^{(c)}, Q_{75}^{(c)}$ as cutoffs, with tiers ordered from below $Q_{25}^{(c)}$ to at or above $Q_{75}^{(c)}$. Summing across the 12 dimensions yields a total in $[0, 36]$ that shares the same range as VInterview-2025, so the level mapping $L(\cdot)$ in Eq. 7 and both reward signals (Eq. 5 and Eq. 6) are reused unchanged. Behavioral indicators $\mathcal{B}(c, s)$ for each of the four tiers are produced by the same MLLM decomposition step (Eq. 3), seeded with the original dimension definitions from (Gupta et al., 2025).

At evaluation time, the model's predicted tiers are mapped back to dimension medians and ranked, so the reported Spearman, Kendall τ-b, C-index, and Pearson scores remain comparable to the original protocol.

C Implementation Details

C.1 Algorithm Details

Algorithm 1 describes the training pipeline of the Scorer π_θ^*, covering supervised fine-tuning followed by GRPO optimization with dual rewards; all other modules in the framework are used zero-shot at inference time. Algorithm 2 details the retrieval, verification, and scoring stages that run for each criterion once the Grapher has built the semantic graph $\mathcal{G}$.

Algorithm 1 Scorer Training

Require: Base model π_0, SFT data $\mathcal{D}_{\text{SFT}}$, rubric set $\{R(c,s)\}$, LLM judge, reward weights λ_1, λ_2, KL coefficient β, group size $|g|$
// Stage 1: Supervised Fine-Tuning
for $(V_j, c_j, s_j^*, r_j^*) \in \mathcal{D}_{\text{SFT}}$ **do**
Fine-tune π_0 on $(V_j, c_j, R(c_j,\cdot)) \to (s_j^*, r_j^*)$, obtaining π_{SFT}
end for
$\pi_{\text{ref}} \leftarrow \pi_{\text{SFT}}$, $\pi_\theta \leftarrow \pi_{\text{SFT}}$
// Stage 2: GRPO with Dual Rewards
for each training step **do**
Sample input (V, C, R) over the full criterion set
Generate group $g = \{(\{\hat{s}_{q,i}\}_{i=1}^N, r_q)\}_{q=1}^{|g|} \sim \pi_\theta$
for each rollout $q \in g$ **do**
$R_{\text{align},q} \leftarrow \text{LLM}_{\text{judge}}(R, \mathcal{B}, \{\hat{s}_{q,i}\}, r_q)$
$\hat{S}_{\text{total},q} \leftarrow \sum_{i=1}^N \hat{s}_{q,i}$
Compute $R_{\text{diff},q}$ via Eq. 6
$R_q \leftarrow \lambda_1 R_{\text{align},q} + \lambda_2 R_{\text{diff},q}$
end for
$\bar{R}_g \leftarrow \text{mean}(\{R_q\})$, $\sigma_g \leftarrow \text{std}(\{R_q\})$
Update π_θ via the group-normalized GRPO objective $\mathcal{L}_{\text{GRPO}}$
end for
return Trained policy π_θ

Algorithm 2 PhoenixNest-Video Inference Pipeline

Require: Video V, criterion c, rubric $R(c,\cdot)$, video graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with prototype entities merged at τ, trained policy π_θ, thresholds θ, δ, retrieval budget N_r
// Step 1: Rubric Decomposition
$\mathcal{B}(c,s) \leftarrow \text{MLLM}(R(c,s))$ for each score level s
// Step 2: Rubric-Conditioned Retrieval
for each behavioral indicator $b_j \in \mathcal{B}(c,s)$ **do**
Generate queries $Q(b_j) \leftarrow \text{MLLM}(b_j)$
$P_j \leftarrow \{e : \max_{q \in Q(b_j)} \text{sim}(q,e) > \theta\}$ {prototypes}
$\mathcal{R}_j \leftarrow \{v_i : v_i \text{ mentions some } e \in P_j\} \cup \{v_i : \text{sim}(Q(b_j), \text{attr}(v_i)) > \theta\}$
end for
$\mathcal{R} \leftarrow \text{Top-}N_r(\bigcup_j \mathcal{R}_j)$ by mean query similarity
// Step 3: Structured Reasoning & Verification
$\mathcal{R}' \leftarrow \emptyset$
for each clip $v_i \in \mathcal{R}$ **do**
for each indicator $b_j \in \mathcal{B}(c,s)$ **do**
// Cross-modal verification, all modalities weighted equally
flags $\leftarrow$ {modality u : $\text{Evidence}_u(v_i, b_j) > \delta$}, $u \in$ {visual, audio, textual}
if flags $\neq \emptyset$ **then**
$\mathcal{R}' \leftarrow \mathcal{R}' \cup \{v_i\}$, record (v_i, b_j, flags)
end if
end for
end for
$A \leftarrow$ recorded entries $(\text{id}(v_i), \text{range}(v_i), b_j, \text{flags})$, in chronological order
// Step 4: Citation-Constrained Generation
$(\hat{s}, r, F) \leftarrow \pi_\theta(V, c, R(c,\cdot), \mathcal{B}(c,\cdot), A)$
Reject if F cites no verified material
return Score $\hat{s}$, rationale r, feedback F, evidence references drawn from $\mathcal{R}'$

C.2 Retrieval Design Ablations

The retrieval stage has two free choices that the main text fixes without justification: the embedding model used for entity matching and the retrieval threshold θ. We ablate both on the VInterview-2025 test set, holding everything else at the default configuration and reporting grade-level accuracy (ACC) as defined in Appendix D.

Embedding Model	ACC ↑
BERT-base-uncased	0.872
CLIP (ViT-L/14) text encoder	0.883
BGE-large-en-v1.5	**0.915**

Table 6: Effect of the retrieval embedding model on VInterview-2025, with $\theta = 0.5$ and $\tau = 0.7$ fixed.

BGE-large-en-v1.5 outperforms both alternatives. BERT-base produces coarser representations that conflate distinct interview topics, merging, for example, "research methodology" with "research motivation." The CLIP text encoder is trained on image captions and transfers poorly to the longer discourse-level descriptions the Grapher extracts from interview clips.

θ	ACC ↑
0.3	0.878
0.4	0.896
0.5	**0.915**
0.6	0.891
0.7	0.874
0.8	0.861

Table 7: Effect of the retrieval threshold θ on VInterview-2025, with BGE-large-en-v1.5 and $\tau = 0.7$ fixed.

Accuracy peaks at $\theta = 0.5$. Below it the Retriever admits loosely related clips that dilute the evidence pool; above it the Retriever misses clips where the candidate demonstrates the target competency with different vocabulary or indirectly, and the Verifier has fewer candidates to work with.

D Metrics Details

For a test set of N_{test} interviews, indexed by n, let $\hat{S}_n$ and S_n denote the predicted and expert total scores for interview n, respectively. Let $\hat{s}_{n,k}$ and $s_{n,k}$ denote the predicted and expert score for criterion $k \in \{1, \ldots, N\}$, where $N = 18$ is the number of rubric criteria; the expert reference is the arithmetic mean of the three faculty ratings and is therefore continuous in $[0, 2]$, and it is rounded to the nearest ordinal level for the contingency-table metrics below. We use $M = 3$ ordinal score levels. The institutional grade mapping is

$$L(S) = \begin{cases} 0, & 0 \le S < 18, \\ 1, & 18 \le S < 24, \\ 2, & 24 \le S < 30, \\ 3, & 30 \le S \le 36. \end{cases} \tag{7}$$

D.1 Total Score Metrics

- **Grade-Level Accuracy (ACC).** Per the participating institution's evaluation guidelines, raw rubric totals are aggregated into the four admissions-relevant tiers in Eq. 7, and faculty reviewers only act on this tier-level outcome rather than on the raw total. We therefore follow this institution-specified scheme and count a prediction as correct when its institutional tier falls within one tier of the expert label (the axis of Figure 4 labels the same quantity within-1-level accuracy):

$$\text{ACC} = \frac{1}{N_{\text{test}}} \sum_{n} \mathbf{1}\big[\,|L(\hat{S}_n) - L(S_n)| \le 1\,\big]. \tag{8}$$

- **Mean Absolute Error (MAE).** The total-score MAE is

$$\text{MAE}_{\text{total}} = \frac{1}{N_{\text{test}}} \sum_{n} \left|\hat{S}_n - S_n\right|. \tag{9}$$

- **Wasserstein Distance (W Dist).** Let $\hat{P}$ and P be the empirical distributions of predicted and expert total scores, with cumulative distribution functions $F_{\hat{P}}$ and F_P. We use the first-order Wasserstein distance:

$$\text{WDist} = \int_{\mathbb{R}} \left|F_{\hat{P}}(x) - F_P(x)\right| dx. \tag{10}$$

Since total scores are discrete in $[0, 36]$, this is equivalently

$$\text{WDist} = \sum_{s=0}^{36} \left|F_{\hat{P}}(s) - F_P(s)\right|. \tag{11}$$

D.2 Per-Criterion Metrics

- **Quadratic Weighted Kappa (QWK).** For each criterion k, let $O^{(k)} \in \mathbb{R}^{M \times M}$ be the observed rating matrix, where $O^{(k)}_{ij}$ counts samples with expert score i and predicted score j. Let

$$E^{(k)}_{ij} = \frac{O^{(k)}_{i\cdot} O^{(k)}_{\cdot j}}{N_{\text{test}}} \tag{12}$$

be the expected count under independent marginals, and let the quadratic weight be

$$w_{ij} = \frac{(i-j)^2}{(M-1)^2}. \tag{13}$$

The per-criterion QWK is then

$$\text{QWK}^{(k)} = 1 - \frac{\sum_{i=0}^{M-1} \sum_{j=0}^{M-1} w_{ij} O^{(k)}_{ij}}{\sum_{i=0}^{M-1} \sum_{j=0}^{M-1} w_{ij} E^{(k)}_{ij}}. \tag{14}$$

The main table reports the macro-average across all criteria:

$$\mathrm{QWK} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{QWK}^{(k)}. \tag{15}$$

- **Per-Criterion MAE.** We first compute the MAE for each criterion and then average across all $N = 18$ criteria:

$$\mathrm{MAE}_{\text{criterion}} = \frac{1}{NN_{\text{test}}}\sum_{k=1}^{N}\sum_{n}\left|\hat{s}_{n,k} - s_{n,k}\right|. \tag{16}$$

D.3 RecruitView Metrics

For the RecruitView benchmark (Gupta et al., 2025), we follow the evaluation protocol of Gupta et al. and report the following four rank-correlation and concordance metrics, macro-averaged over all $T = 12$ regression targets (the Big Five personality traits, an overall-personality index, and six interview-performance dimensions). Let $\hat{y}_n$ and y_n denote the predicted and annotated scores for sample n, respectively.

- **Spearman Rank Correlation (ρ).** Measures the monotonic association between predicted and annotated rankings. Let $\mathrm{rk}(\cdot)$ denote the rank operator:

$$\rho = 1 - \frac{6\sum_{n=1}^{N}\big(\mathrm{rk}(\hat{y}_n) - \mathrm{rk}(y_n)\big)^2}{N(N^2-1)}. \tag{17}$$

- **Kendall Rank Correlation (τ-b).** Counts concordant and discordant pairs, with a correction for ties:

$$\tau_b = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}, \tag{18}$$

where n_c and n_d are the numbers of concordant and discordant pairs, $n_0 = \binom{N}{2}$, n_1 is the number of pairs tied in $\hat{y}$, and n_2 is the number of pairs tied in y.

- **Concordance Index (C-index).** The fraction of all comparable pairs whose predicted ordering agrees with the annotated ordering. Let

$$N_c = \#\{(i,j) : i<j,\ (\hat{y}_i - \hat{y}_j)(y_i - y_j) > 0\},$$
$$N_t = \#\{(i,j) : i<j,\ \hat{y}_i = \hat{y}_j\},$$
$$N_p = \#\{(i,j) : i<j,\ y_i \neq y_j\}.$$

Then

$$\text{C-index} = \frac{N_c + \frac{1}{2}N_t}{N_p}. \tag{19}$$

- **Pearson Correlation (r).** The linear correlation between predicted and annotated scores of each target dimension:

$$r = \frac{\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})(y_n - \bar{y})}{\sqrt{\sum_{n=1}^{N}(\hat{y}_n - \bar{\hat{y}})^2}\sqrt{\sum_{n=1}^{N}(y_n - \bar{y})^2}}. \tag{20}$$

All four metrics are computed per target and then macro-averaged:

$$\overline{M} = \frac{1}{T}\sum_{t=1}^{T} M^{(t)}, \quad M \in \{\rho, \tau_b, \text{C-index}, r\}. \tag{21}$$

E Baselines

To evaluate the performance of our approach, we compare against the following methods:

- **GPT-5.4** (Singh et al., 2025): A frontier proprietary multimodal model with strong reasoning capabilities. We evaluate it in a zero-shot setting with rubric descriptors.
- **Claude-opus-4-6** (Anthropic, 2025): A proprietary large language model known for nuanced reasoning and instruction following. We use it as a zero-shot evaluator for video interview assessment.
- **Gemini-3.1-pro-preview** (Team et al., 2023): A proprietary multimodal model from Google with long-context understanding capabilities. We evaluate it under the same prompt-based evaluation protocol.
- **Grok-4.1** (xAI, 2025): A frontier proprietary model optimized for fast inference. We evaluate it under the same zero-shot prompt-based protocol.
- **Doubao-Seed-1.8** (Guo et al., 2025): A proprietary multimodal model evaluated on RecruitView only, under the same zero-shot prompt-based protocol.
- **Qwen3.5-397B-A17B** (Bai et al., 2025): A large open-source mixture-of-experts multimodal model. We evaluate it zero-shot with rubric-conditioned prompts.
- **Qwen3.5-27B** (Bai et al., 2025): A mid-scale open-source multimodal model from the same family. We evaluate it under the identical prompt-based protocol.
- **Kimi-K2.5** (Team et al., 2026): An open-source long-context multimodal model designed for extended-document and video reasoning. We evaluate it zero-shot with rubric descriptors.

- **GLM-4.5v** (Hong et al., 2025): An open-source multimodal model with strong vision-language alignment. We apply it to interview assessment using the same prompt template.
- **Video-R1** (Feng et al., 2025): A video understanding model that applies reinforcement learning to elicit chain-of-thought reasoning in video contexts. We evaluate it with rubric descriptors for structured assessment.
- **VideoChat-R1** (Li et al., 2025b): A video model enhanced with reinforcement fine-tuning for spatio-temporal perception. We use it as a baseline under the same prompt-based evaluation protocol.
- **VideoRFT** (Wang et al., 2025a): A video model trained with reinforced fine-tuning on video reasoning tasks. We apply it to interview evaluation under the same protocol.
- **VideoLLaMA2** (Cheng et al., 2024): A video-language model with improved spatio-temporal modeling and audio understanding. We use it as a video-specific baseline with the standard evaluation prompt.
- **VideoLLaMA3** (Zhang et al., 2025): A frontier multimodal foundation model for both image and video understanding. We evaluate it using the same rubric-conditioned prompt approach.

F Rubric Definitions

Tables 8 and 9 present the complete scoring rubric used in our video interview evaluation. Each criterion is scored on a three-level scale: 0 (Needs Improvement), 1 (Meets Expectations), and 2 (Exceeds Expectations). The total score ranges from 0 to 36.

Overall Scoring Summary. The total score is the sum of all 18 criteria ($S_{total} \in [0, 36]$):

- $S_{total} \in [30, 36]$: Outstanding potential with excellent personal qualities.
- $S_{total} \in [24, 30)$: Good potential with strong performance across both phases, but some room for improvement remains.
- $S_{total} \in [18, 24)$: Average potential with deficiencies in key areas like communication and teamwork.
- $S_{total} \in [0, 18)$: Limited potential with significant shortcomings or lack of standout performances.

G Additional Figures

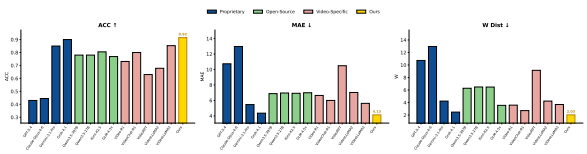


Figure 6: Bar chart comparison of main evaluation metrics across all baselines on the VInterview-2025 test set. PhoenixNest-Video achieves the highest grade-level accuracy and the lowest total-score MAE and Wasserstein distance among all methods.

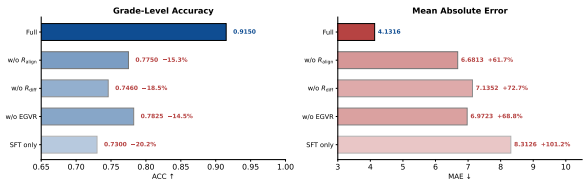


Figure 7: Ablation study results visualized as a bar chart. Each bar group corresponds to a model variant with one component removed, illustrating the individual contribution of the alignment reward, differentiation reward, and retrieval-and-verification pipeline.

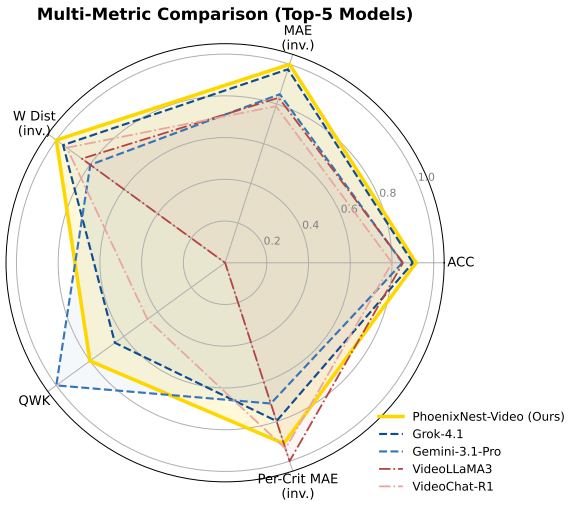


Figure 8: Radar chart comparing the top five performing models across all evaluation dimensions. Each axis represents a distinct metric, highlighting the relative strengths and weaknesses of each model.

H Core Prompts and Schemas

We list the prompts used at inference time by the Grapher, Retriever, and Verifier. The Scorer is not prompted zero-shot: it is the trained policy π_θ^* and receives the rubric, the behavioral indicators, and the reasoning chain A in the structured format established during supervised fine-tuning.

H.1 Entity Extraction (Grapher)

This prompt is applied to each K-frame clip together with its aligned transcript segment. Its output defines the nodes and node attributes of the video graph.

```
Analyze the given interview video clip and
extract key information in a structured JSON
format in English. Identify and describe:

Entities: List all distinct people, objects,
presentation materials, or other significant
elements visible in the clip (e.g., the
candidate, slides, diagrams, demonstration
materials).

Actions: Describe the actions, behaviors, and
interactions of the identified entities (e.g.,
presenting slides, responding to a panel
question, explaining a research methodology,
demonstrating a prototype, making eye contact
with the panel).

Scenes: Identify the interview context or phase
(e.g., presentation phase, Q&A interaction,
slide transition, panel discussion).

Ensure the output strictly follows the JSON
format below:
{
    "entities": [
        {"entity name": "", "description": ""}
    ],
    "actions": [
        {"entity name": "",
         "action description": ""}
    ],
    "scenes": [
        {"location": ""}
    ]
}

The "entity name" in actions must correspond to
an "entity name" in entities. Each section
should be detailed but concise, capturing all
relevant interactions and contextual elements
from the clip. Output only the JSON object.
```

Each `entities[].entity name` is a short label such as "candidate" or "research poster," and the paired `description` is the clip-local context that is embedded for entity merging and retrieval. Each `actions[].entity name` must refer back to a declared entity, and `scenes[].location` carries an interview-phase label rather than a physical place.

H.2 Rubric Decomposition

For each criterion c and score level s, this prompt expands the rubric descriptor $R(c, s)$ into the behavioral indicators $\mathcal{B}(c, s)$ of Eq. 3.

```
Given the following rubric criterion and
score-level descriptor, generate a list of
concrete, observable behavioral indicators that
a candidate would demonstrate during an
interview to receive this score level.

Criterion: {criterion_name}
Score Level: {score_level}
Descriptor: {rubric_descriptor}

Output a JSON list of behavioral indicators.
Each indicator should describe a specific,
observable behavior that can be verified from
video, audio, or transcript evidence.

Output format:
[
    "indicator 1: ...",
    "indicator 2: ...",
    ...
]
```

For Critical Thinking at $s = 2$, the decomposition yields indicators such as "analyzes the problem from multiple complementary angles," "supports claims with specific evidence from the literature or data," and "identifies limitations of the proposed approach and addresses them." At $s = 0$ it yields "provides circular or unsubstantiated reasoning" and "fails to engage with the substance of the question."

H.3 Retrieval Query Refinement (Retriever)

Each behavioral indicator b_j is rewritten into targeted queries $Q(b_j)$ that are matched against the entity index of the video graph.

```
Given the following behavioral indicator from an
interview assessment rubric, generate targeted
search queries to locate relevant video clips in
an interview recording.

Behavioral indicator: {indicator}
Criterion: {criterion_name}

Generate 2-3 concise search queries that capture
the core semantic content of this indicator.
Each query should focus on observable actions,
topics, or behaviors that would appear in the
video, audio, or transcript.

Output format:
["query 1", "query 2", ...]
```

H.4 Verification (Verifier)

For each candidate clip v_i and indicator b_j, the Verifier judges every modality independently, so that the three channels enter the decision with equal weight.

```
You are given a video clip from a candidate
interview and a behavioral indicator from the
assessment rubric.

Behavioral indicator: {indicator}

Judge each modality separately. For the visual
channel consider the candidate's actions, body
language, and slides; for the audio channel the
delivery and the transcribed speech; for the
```

```
textual channel the transcript and the written
presentation materials.

For each channel, state whether it supports the
indicator and give a confidence in [0, 1].

Output format:
{"visual":  {"supported": true, "conf": 0.0},
 "audio":   {"supported": true, "conf": 0.0},
 "textual": {"supported": true, "conf": 0.0},
 "evidence": "brief description"}
```

A channel whose confidence exceeds δ sets the corresponding verification flag in the reasoning chain A, and a clip is retained when at least one flag is set. The `evidence` field supplies the human-readable justification that the Scorer cites in its final report.

Table 8: Phase I: Presentation Rubric (Max 18 Points)

Criterion	**0 – Needs Improvement**	**1 – Meets Expectations**	**2 – Exceeds Expectations**
c_1: Topic Relevance	Speech was not focused on the chosen topic or deviated significantly from the designated topics.	Speech was related to the topic and covered key aspects, but was presented in a straightforward manner.	Speech was tightly focused on the topic, problem-oriented or narrative-driven, with clear thematic coherence throughout.
c_2: Perspective & Vision	No multi-angle thinking demonstrated; did not address real-world problems or future challenges.	Showed some multi-angle thinking or addressed a real-world problem at a surface level.	Demonstrated rich multi-angle perspectives, addressing real-world problems or future challenges with depth and insight.
c_3: Organization & Structure	Speech lacked clear structure; no identifiable beginning, body, or conclusion.	Speech had a basic structure with recognizable sections and some use of transitional language.	Speech was well-organized with clear segments, effective transitions, and a cohesive narrative flow from opening to conclusion.
c_4: Content Familiarity	Frequent pauses, reading from slides or notes, showing limited understanding of the content.	Accurately explained concepts and backgrounds without reading, demonstrating adequate understanding.	Spoke fluently with deep understanding, simplifying complex topics and enriching the presentation with professional knowledge or examples.
c_5: Slide Design	Slides contained errors, were inconsistent with speech content, or had excessive text and poor design.	Slides were generally consistent with the speech content, free of errors, and balanced in their use of text and visuals.	Slides were well-designed, accurate, visually appealing, and effectively supported the speech with appropriate illustrations.
c_6: Language Expression	Expression was frequently interrupted, unclear, with excessive filler words and vague language.	Expression was generally fluent and clear, with only occasional filler words.	Expression was highly fluent, articulate, and precise, with no unnecessary filler words and clear, confident delivery.
c_7: Engagement with Judges	No meaningful eye contact or interaction with judges during the presentation.	Basic eye contact and some interaction with judges during the presentation.	Effective and sustained eye contact with judges, actively engaging them throughout the presentation with confident body language.
c_8: Creative Thinking	No creative or unique perspectives were offered.	Proposed some creative or cross-disciplinary ideas.	Proposed creative or unique perspectives with strong persuasiveness and supporting evidence.
c_9: Motivation & Fit	Did not explain application motivation or suitability for the program.	Adequately explained application motivation and general suitability for the program.	Clearly articulated personal goals aligned with program objectives, demonstrating strong understanding of the program and compelling reasons for the fit.

Table 9: Phase II: Question & Answer Rubric (Max 18 Points)

Criterion	**0 – Needs Improvement**	**1 – Meets Expectations**	**2 – Exceeds Expectations**
c_{10}: Comprehension Ability	Did not understand or accurately respond to the judges' questions.	Basically understood and responded to the judges' questions.	Carefully listened to and accurately understood the judges' questions, providing precise and targeted responses.
c_{11}: Answer Depth	Answers were superficial, showing limited understanding of the topic.	Answers were adequate and showed reasonable understanding after some thought.	Answers were well-thought-out, demonstrating deep understanding with supporting examples, data, or professional insights.
c_{12}: Emotional Regulation	Showed emotional instability, anxiety, or loss of composure under pressure.	Maintained basic composure and stable emotional state during challenging questions.	Remained calm, confident, and composed under pressure, demonstrating resilience and positive demeanor throughout.
c_{13}: Crisis Management	When facing challenging or unexpected questions, responses were vague, evasive, or off-topic.	Gave reasonable responses to challenging questions, proposing some solutions or ideas.	Demonstrated flexible and quick thinking, immediately providing clear, well-organized solutions with high persuasiveness.
c_{14}: Critical Thinking	Answers lacked logical reasoning or critical analysis.	Answers showed some logical reasoning and basic analytical ability.	Answers demonstrated clear causal reasoning, multi-angle analysis, and insightful critical thinking with well-supported arguments.
c_{15}: Teamwork Focus	No mention of teamwork experiences or collaborative activities.	Mentioned teamwork keywords and provided some examples of collaborative experiences.	Provided specific, detailed examples of teamwork demonstrating collaboration skills and strong interest in collective achievement.
c_{16}: Persuasiveness	Arguments lacked conviction; unable to effectively support their position.	Arguments were generally convincing with basic supporting evidence.	Arguments were highly persuasive, well-supported with compelling evidence, and effectively defended their position.
c_{17}: Self-Awareness & Growth	Did not discuss personal strengths, weaknesses, or areas for improvement.	Basically analyzed personal strengths and weaknesses.	Demonstrated deep self-awareness, thoughtfully discussing strengths, weaknesses, and specific directions for improvement with concrete plans.
c_{18}: Empathy & Engagement	Ignored the judges' reactions and feedback during Q&A.	Made basic adjustments based on the judges' reactions and feedback.	Carefully observed the judges' reactions, proactively adjusted responses, and engaged in meaningful dialogue.