MIS-Bench: Benchmarking Multimodal LLMs for Psychotherapeutic Interpersonal Skills Assessment

Yuhan Lu, Yi Yao, Hua Shen*, Katie Aafjes-van Doorn*, Zhaonan Wang*
NYU Shanghai, New York University
{yl12006, yy5074, huashen, kav9239, zhaonan.wang}@nyu.edu

Abstract

Multimodal large language models (MLLMs) are increasingly used as evaluators, yet their reliability in professional assessment tasks that require expert judgment remains unclear. We investigate this challenge in the context of assessing psychotherapeutic interpersonal skills and introduce **MIS-Bench**, a Multimodal Interpersonal Skills (MIS) benchmark comprising 996 psychotherapy response videos annotated across 8 dimensions of Facilitative Interpersonal Skills. Across 9 MLLMs with multiple modality and prompting settings, we find that current models show only modest agreement with human experts, inconsistent gains from multimodal input, and limited benefits from reasoning-based prompting. To mitigate this gap, we propose **MIS-RAFT**, a regression-aware fine-tuning method inspired by RAFT and tailored to fine-grained interpersonal skill scoring at one-decimal precision. MIS-RAFT addresses the mismatch between autoregressive token prediction and scalar-valued expert assessment, significantly improving agreement with human ratings. Overall, MIS-Bench reveals a clear gap between general multimodal capability and expert-level interpersonal judgment, while MIS-RAFT offers a promising path toward more reliable model-based assessment.[1]

1 Introduction

Multimodal large language models (MLLMs) are increasingly used as evaluators (Zhu et al., 2023; Thakur et al., 2024; Li et al., 2024), but their reliability in **expert-based professional assessment** remains underexplored. Unlike general multimodal tasks, professional assessment often requires calibrated judgments over subtle human behaviors, domain-specific rubrics, and expert scores. Psychotherapeutic interpersonal skill assessment provides a representative and challenging setting, as accurate evaluation requires integrating what a therapist says with how the response is delivered through paralinguistic and nonverbal cues.

We study this problem through the **Facilitative Interpersonal Skills (FIS)** framework, a standardized performance-based measure of therapist interpersonal competence. In FIS, clinicians respond to challenging patient scenarios, and trained human raters score their responses across eight dimensions, including verbal fluency, emotional expression, warmth, empathy, alliance bond capacity, and alliance rupture-repair responsiveness (Anderson et al., 2009, 2016b). These dimensions were carefully developed and clinically grounded, making FIS a rigorous testbed for evaluating how well MLLMs can approximate expert psychotherapist judgment.

In this work, we introduce **MIS-Bench**, a Multimodal Interpersonal Skills benchmark for evaluating MLLMs on psychotherapeutic interpersonal skill assessment, as illustrated in Figure 1. MIS-Bench contains 996 psychotherapy response videos annotated across the eight FIS dimensions, with transcript information, rubric descriptions, and expert ratings. Compared with existing multimodal benchmarks that emphasize general perception, question answering, or affect recognition (Chen et al., 2024; Hu et al., 2025; Zhang et al., 2025), MIS-Bench introduces several key advances: ❶ **Inherently Multimodal Evidence**, as human raters are also instructed to attend not only to verbal content but also to delivery, nonverbal cues; ❷ **Fine-grained Multidimensional Scoring**, organizing evaluation around eight FIS dimensions to enable more nuanced and rigorous evaluation; ❸ **Expert-Based Professional Assessment**, focusing on interpersonal skills evaluation in clinical psychotherapy, MIS-Bench provides a useful probe of whether MLLMs trained primarily on large-scale web data can generalize to specialized professional evalua-

*Corresponding authors.

[1]Data and code are released on GitHub: https://github.com/calistalu/MIS-Bench.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 31582–31607
October 24-29, 2026 ©2026 Association for Computational Linguistics

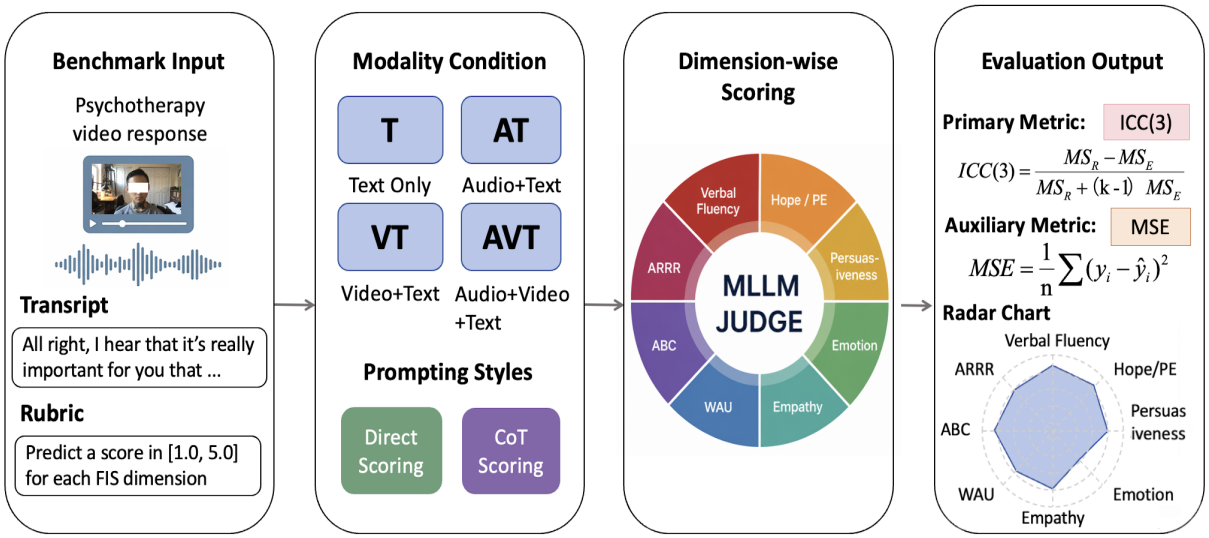


Figure 1: **Overview of MIS-Bench.** Each benchmark instance is built from a psychotherapy response video and associated transcript and rubric information. Models are evaluated with 4 modality configurations and 2 prompting styles, and predict one scalar score for each FIS dimension. Performance is measured by ICC(3) and MSE.

tion.

In MIS-Bench, nine open-source and proprietary MLLMs with 4 modality configurations and 2 prompting strategies are being evaluated. Our results show that current models have only modest agreement with human experts, and performance gains vary substantially across dimensions and inconsistently across modality configurations. We also find that reasoning-based prompting offers limited benefits compared with direct scoring, suggesting that eliciting rationales does not necessarily lead to better-calibrated professional judgments.

To mitigate this gap, we further propose **MIS-RAFT**, a regression-aware fine-tuning method for multimodal interpersonal skill scoring. Inspired by Regression-Aware Fine-Tuning (Lukasik et al., 2025), MIS-RAFT adapts the objective to fine-grained FIS score prediction at one-decimal precision. Rather than treating scores as autoregressive tokens, MIS-RAFT directly optimizes scalar expert-score prediction, thereby bridging the mismatch between token generation and numerical professional assessment. Experiments show that MIS-RAFT substantially improves agreement with human ratings. In summary, our contributions are threefold:

- We introduce **MIS-Bench**, a benchmark comprising 996 psychotherapy response videos for multimodal, expert-based interpersonal skill assessment across eight FIS dimensions.
- We systematically evaluate nine MLLMs across various modality and prompting settings, revealing modest agreement with human experts, inconsistent gains from multimodal input, and limited benefits from reasoning-based prompting.
- We propose **MIS-RAFT**, a regression-aware fine-tuning method that mitigates the objective mismatch between autoregressive token prediction and scalar-valued expert scoring, significantly improving agreement with human ratings.

2 Related Work

2.1 MLLMs as Evaluators

Large language models have increasingly been used as automated evaluators for open-ended generation, giving rise to the LLM-as-a-Judge paradigm (Gu et al., 2026; Zhu et al., 2023). More recent work extends this idea to multimodal settings. In particular, Chen et al. (2024) provide one of the first systematic studies of MLLMs as judges in vision-language evaluation, while Ko et al. (2025) show that text-only reasoning supervision can improve zero-shot multimodal evaluator performance. Most existing work, however, centers on general-purpose tasks such as pairwise comparison, ranking, and discrete scoring. MIS-Bench focuses on fine-grained assessment in a specialized professional domain, where the goal is to align with expert human ratings of psychotherapeutic interpersonal skills rather than to evaluate generic multimodal outputs.

2.2 Existing Multimodal Benchmarks for Human-Centered Assessment

A growing line of work studies multimodal benchmarks for human-centered capabilities such as emotion understanding and interpersonal skill assessment. Recent benchmarks including EmoBench-M (Hu et al., 2025) and MME-Emotion (Zhang et al., 2025) evaluate emotional intelligence in MLLMs across diverse multimodal scenarios. Broader reviews have also examined automatic multimodal interpersonal skill assessment in social and educational settings (Rasipuram and Jayagopi, 2020; Guerrero-Sosa et al., 2025). MIS-Bench addresses a different target: rather than general affect understanding or broad interpersonal behavior, it targets professional psychotherapy skill assessment grounded in a clinically established framework and evaluated against expert ratings.

2.3 Computational Assessment in Psychotherapy

Within psychotherapy research, the FIS framework has been studied as a clinically meaningful, performance-based measure of therapists' interpersonal skill, with prior reviews summarizing its theoretical basis, psychometric properties, and potential for wider use (Allen et al., 2024; Anderson et al., 2009, 2016b,a, 2020). Attempts to automate FIS scoring have so far relied on conventional supervised models rather than MLLMs. Early efforts used conventional machine learning to predict FIS scores from linguistic or multimodal features (Goldberg et al., 2021; Zech et al., 2022), and more recent work has proposed a multimodal proof of concept for AI-based FIS measurement (Aafjes-van Doorn et al., 2025), alongside discussions of its future potential for scalable assessment (Aafjes-van Doorn and Lin, 2024). Our work builds on this foundation, but shifts the focus from building a single automated scorer to systematically benchmarking modern open-source and closed-source MLLMs on this expert-grounded task.

3 MIS-Bench

3.1 Task Definition

Grounded in the theory of *Facilitative Interpersonal Skills* (FIS), MIS-Bench evaluates whether an MLLM can approximate the domain-expert assessment of psychotherapeutic interpersonal skills on a calibrated scale. Let $\mathcal{D} = \{d_1, \ldots, d_8\}$ denote the eight FIS dimensions: *Verbal Fluency (VF)*, *Hope and Positive Expectations (H/PE)*, *Persuasiveness (Per.)*, *Emotional Expression (Emo.)*, *Warmth, Acceptance, and Understanding (WAU)*, *Empathy (Emp.)*, *Alliance Bond Capacity (ABC)*, and *Alliance Rupture–Repair Responsiveness (ARRR)* (Anderson et al., 2009; Allen et al., 2024). For each clinician response instance i, we denote its video, audio, and transcript as V_i, A_i, and T_i, respectively. Each dimension $d \in \mathcal{D}$ is then associated with a rubric description R_d and an expert-aggregated human score $y_{i,d} \in [1.0, 5.0]$ obtained from trained human observer ratings.

Table 1: Released feature structure of MIS-Bench. The public release contains de-identified feature representations rather than raw audiovisual recordings.

Modality	Feature	Shape / Type
Face	Action units	$T \times 8$
Face	Facial landmarks (98 points, x/y)	$T \times 196$
Face	Gaze direction	$T \times 2$
Face	Detection confidence	$T \times 2$
Audio	Wav2Vec 2.0 frame embeddings	$T \times 768$
Audio	Prosody	$T \times 15$
Audio	Log-Mel spectrogram	$T \times 80$
Audio	MFCCs	$T \times 13$
Text	Transcript words	N_{word}
Label	FIS subscale scores	8 scores in $[1.0, 5.0]$

We formulate MIS-Bench as a dimension-wise scoring task. Specifically, each clinician-response and dimension pair (i, d) is treated as an independent evaluation instance. Given a modality configuration $m \in \{\text{text-only}, \text{audio+text}, \text{video+text}, \text{audio+video+text}\}$, we define the corresponding model input as X_i^m, which contains the available modalities, transcript text, and rubric information under that setting. Under prompting strategy $s \in \{\text{direct}, \text{CoT}\}$, an MLLM f_θ produces a response

$$A_{i,d}^{m,s} = f_\theta(P_s, X_i^m, R_d), \tag{1}$$

where P_s denotes the prompt template, and $A_{i,d}^{m,s}$ is either a direct numeric answer or a reasoning trace followed by a final numeric score. We then extract the predicted scalar score using a deterministic parser $g(\cdot)$:

$$\hat{y}_{i,d}^{m,s} = g(A_{i,d}^{m,s}), \quad \hat{y}_{i,d}^{m,s} \in [1.0, 5.0]. \tag{2}$$

Table 2: Human single-rater agreement ceiling on the 625 fully rated videos. Values are single-measure ICC(3,1) scores reported in percentage points; brackets denote 95% confidence intervals. Higher is better.

	VF	H/PE	Per.	Emo.	WAU	Emp.	ABC	ARRR
Human ICC(3,1)	64.0	53.0	54.0	51.0	50.0	43.0	51.0	53.0
95% CI	[61.0, 68.0]	[49.0, 58.0]	[49.0, 58.0]	[46.0, 55.0]	[45.0, 54.0]	[38.0, 48.0]	[46.0, 55.0]	[48.0, 57.0]

The benchmark evaluates how closely $\hat{y}_{i,d}^{m,s}$ approximates the expert-aggregated human score $y_{i,d}$ across dimensions, models, modalities, and prompting strategies.

3.2 Benchmark Construction

The current release of MIS-Bench contains 996 clinician response videos annotated under the FIS framework. Each benchmark instance is derived from a recorded clinician response video and includes associated metadata, response text fields, score annotations, and rubric descriptions, and human ratings for the eight FIS dimensions. To protect participant privacy, the public release does not include raw audiovisual recordings. Instead, it provides de-identified multimodal representations, transcripts, and FIS scores. Table 1 summarizes the released feature structure.

MIS-Bench is constructed by repurposing clinician responses collected under the original FIS performance task into a benchmark for multimodal model evaluation. In the source protocol, clinicians first view standardized patient stimulus clips portraying a challenging therapeutic encounter (e.g., the patient wants to stop therapy), and then provide an immediate spoken response as if they were in session with the patient (Anderson et al., 2009, 2020). The ground truth labels in MIS-Bench are derived from trained human observer-raters following the original FIS rating manual. Raters are instructed to consider both the *content* of the response (verbal information) and its *delivery* (nonverbal and paraverbal information) when assigning scores. Each response is evaluated on the eight FIS subdimensions using a 5-point rating scale from 1 (*skill deficit*) to 5 (*optimal demonstration of skill*), where 3 corresponds to the ordinary level of FIS expected of a trained clinician (Anderson et al., 2009).

Each video was independently rated by trained expert raters following the FIS manual. For benchmark use, we aggregate available ratings for each video–dimension pair by taking their arithmetic mean and retaining one decimal place, which serves as the ground-truth target in all experiments. Of the 996 videos, 625 have complete ratings from all three assigned raters across all eight dimensions. Table 2 reports single-measure ICC(3,1) of human agreement ceiling and 95% confidence intervals among the human raters.

3.3 Evaluation Metrics

We use **ICC(3)** as the primary evaluation metric. ICC(3) directly measures agreement between the model and human raters while treating the model as a fixed additional rater. Higher ICC(3) values indicate stronger agreement with human expert judgment.

For each FIS dimension, let y_{ij} denote the score assigned to response instance i by rater j, where the raters correspond to the expert-aggregated human score and the model prediction. Let n be the number of response instances and k be the number of raters. We compute ICC(3,1) using the two-way mixed-effects formulation:

$$\mathrm{ICC}(3,1) = \frac{MS_R - MS_{\mathrm{err}}}{MS_R + (k-1)MS_{\mathrm{err}}}, \quad (3)$$

where MS_R is the mean square for response instances and MS_{err} is the residual mean square:

$$MS_R = \frac{k}{n-1}\sum_{i=1}^{n}(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2, \quad (4)$$

$$MS_{\mathrm{err}} = \frac{1}{(n-1)(k-1)} \times \sum_{i=1}^{n}\sum_{j=1}^{k}\left(y_{ij} - \bar{y}_{i\cdot} - \bar{y}_{\cdot j} + \bar{y}_{\cdot\cdot}\right)^2. \quad (5)$$

Here, $\bar{y}_{i\cdot}$ is the mean score for response instance i across raters, $\bar{y}_{\cdot j}$ is the mean score assigned by rater j across response instances, and $\bar{y}_{\cdot\cdot}$ is the grand mean.

As an auxiliary error metric, we report MSE in Appendix D.2 to measure the absolute numerical deviation between the model-predicted score and the expert-aggregated human score. Lower MSE indicates better calibration to the human expert rating.

Table 3: **Overall performance comparison on MIS-Bench using dimension-level ICC (%).** For each model, we report the dimension-level ICC values from the configuration that achieves the highest overall score. This overall score is computed by averaging the eight predicted dimension scores for each video and comparing the resulting average against the corresponding average of the human ratings. If a model output is unparseable for any of the eight dimensions, that video is excluded when determining the overall score, with a mean exclusion rate of 1.5%. The top three performing results in each dimension are highlighted in red (1st), blue (2nd), and yellow (3rd) backgrounds, respectively.

Model	LLM Size	A	V	T	VF	H/PE	Per.	Emo.	WAU	Emp.	ABC	ARRR
Open-source MLLMs												
Qwen2-Audio-7B-Instruct (Chu et al., 2024)	7B	✓		✓	-6.0	17.1	8.4	0.9	8.4	2.3	9.1	0.1
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	7B		✓	✓	2.8	39.9	29.9	16.5	36.4	28.8	33.0	27.4
Qwen2.5-Omni-7B (Xu et al., 2025a)	7B	✓	✓	✓	-3.1	19.4	15.3	11.0	18.6	23.0	26.1	20.1
Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b)	30B	✓	✓	✓	21.2	24.6	25.1	14.8	22.8	24.6	18.6	22.7
InternVL3.5-8B-HF (Wang et al., 2025)	8B		✓	✓	5.9	32.4	16.6	14.7	28.9	22.9	26.8	17.6
MiniCPM-V-4.5 (Yu et al., 2025)	9B		✓	✓	8.8	26.9	17.2	14.3	25.1	22.6	32.8	18.5
MiniCPM-o-4.5 (OpenBMB, 2026)	9B	✓	✓	✓	30.5	32.6	26.7	21.9	32.4	29.2	31.0	22.1
Closed-source MLLMs												
GPT-5.4 (OpenAI, 2026)	—		✓	✓	40.1	46.4	31.2	21.8	46.7	42.4	36.5	43.3
Gemini-3.1-Pro-Preview (Google, 2026)	—	✓	✓	✓	17.5	37.8	35.9	32.0	33.9	25.3	38.7	46.6

4 Experiments

4.1 Experimental Setup

Model Selection. We evaluate the performance on a total of nine cutting-edge MLLMs, including Qwen2-Audio-7B-Instruct (Chu et al., 2024), Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen2.5-Omni-7B (Xu et al., 2025a), Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025b), InternVL3.5-8B-HF (Wang et al., 2025), MiniCPM-V-4.5 (Yu et al., 2025), MiniCPM-o-4.5 (OpenBMB, 2026), GPT-5.4 (OpenAI, 2026), and Gemini-3.1-Pro-Preview (Google, 2026). For most open-sourced MLLMs, we use the vLLM framework (Kwon et al., 2023) to perform inference on our benchmark. For models that are not supported by vLLM or require model-specific loading and preprocessing, we instead use a unified evaluation pipeline built on the Hugging Face Transformers library (Wolf et al., 2020), following the official open-source code and model cards provided by the authors. For closed-sourced MLLMs, inference is conducted directly via their official APIs. The main benchmark is conducted under zero-shot settings. To ensure a fair comparison across models, we maintain consistent prompting wherever possible, with only minor modifications when necessary. For example, for audio-language models, the word "video" in prompts is replaced with "audio" when the input modality is audio only. Beyond the main zero-shot evaluation, we further conduct a targeted few-shot CoT analysis on MiniCPM-o-4.5.

Evaluation Strategy. We evaluate each model under four modality configurations: text-only, audio+text, video+text, and audio+video+text. The text-only setting provides only the transcript; audio+text provides the audio signal together with the transcript; video+text provides the video signal together with the transcript; and audio+video+text provides visual, auditory, and textual inputs. Since not all models support all input modalities, each model is evaluated only under the modalities supported by its interface. For each modality configuration, we compare two prompting strategies: direct scoring and chain-of-thought (CoT) prompting. In the direct setting, the model is instructed to output only a numeric FIS score. In the CoT setting, the model is asked to first provide an explicit analysis of the observed therapist behaviors and then produce a final numeric score. The full prompt templates for all eight benchmark settings are provided in Appendix A.1.

4.2 Main Results

We present the overall benchmark comparison in Table 3, and the fine-grained dimension-level results in Figure 2. Averaging the eight dimension-level ICC values, the top-performing models are GPT-5.4 (OpenAI, 2026), Gemini-3.1-Pro-Preview (Google, 2026), and MiniCPM-o-

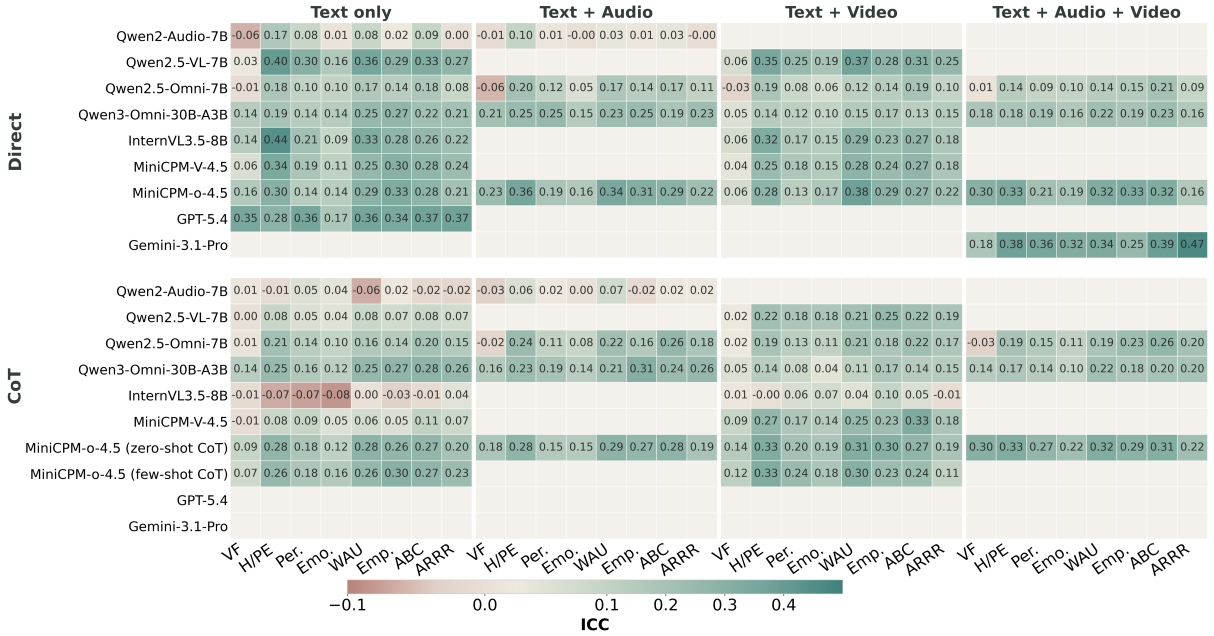


Figure 2: **Dimension-level ICC heatmaps on MIS-Bench.** We report every supported model × modality × prompting combination across four modality configurations and two prompting strategies. In each panel, rows correspond to models and columns correspond to the eight FIS dimensions. Higher ICC indicates better agreement with human ratings, and blank cells indicate configurations unsupported by a given model. The complete grid enables controlled comparisons among models within a common configuration.

4.5 (OpenBMB, 2026), with average ICC scores of 38.6%, 33.5%, and 28.3%, respectively. Human single-rater agreement provides an empirical context for interpreting these model results: human ICC ranges from 43.0% to 64.0% across 8 dimensions (Table 2). In comparison, all evaluated MLLMs report average dimension-level ICC scores below 40%, and all open-source MLLMs remain below 30%, well below the human agreement ceiling. This performance landscape highlights the challenging nature of the MIS-Bench benchmark. It also suggests that even state-of-the-art MLLMs are still in the early stages of developing reliable expert-level judgment for psychotherapy skill assessment.

Among open-source models, no single model family dominates across all settings. Qwen2.5-VL-7B-Instruct performs strongly under text-only direct prompting, Qwen3-Omni-30B-A3B-Instruct reaches its best result under audio+text direct prompting, and MiniCPM-o-4.5 performs best under video+text direct prompting. This pattern suggests that strong performance on MIS-Bench can arise from different model design choices, but it also indicates that current systems are highly sensitive to input configuration and prompting strategy.

Figure 2 further reveals substantial variation across FIS dimensions. Some dimensions are consistently easier than others: averaged across evaluated configurations, *Hope and Positive Expectations* achieves the highest mean ICC (0.214), followed by *Alliance Bond Capacity* (0.206) and *Warmth, Acceptance, and Understanding* (0.206). In contrast, *Emotional Expression* (0.108) and especially *Verbal Fluency* (0.071) are much more difficult. This gap suggests that current MLLMs are relatively better at capturing broad relational and supportive qualities than fine-grained delivery-related skills that depend on subtle emotional or acoustic cues.

4.3 Observations and Insights

Delving into the experimental results, we derive the following key observations and insights.

Obs. 1: Current MLLMs remain limited on expert-grounded psychotherapy skill assessment. Although several models achieve non-trivial alignment with human raters on this expert-grounded task, even the strongest results remain a meaningful gap. And performance varies substantially across models and settings. This result confirms that MIS-Bench is a difficult test of domain-specific judgment requiring nuanced interpretation of therapist behavior. More broadly, it suggests that

strong general-purpose multimodal capability does not automatically translate into reliable evaluation of interpersonal skills in psychotherapy settings.

Obs. 2: Direct score prediction is generally more reliable than CoT prompting. Across matched model–modality comparisons, direct prompting achieves an average overall ICC improvement of 0.047 over zero-shot CoT prompting across 20 matched pairs. At the dimension level, this corresponds to 160 matched comparisons across eight FIS dimensions, where direct prompting achieves higher ICC in 96/160 comparisons and lower MSE in 159/160 comparisons. We also evaluate few-shot CoT on MiniCPM-o-4.5 in two representative settings: text-only and video+text. Although few-shot CoT improves over zero-shot CoT, it still achieves lower ICC compared with direct prompting in 14/16 corresponding dimension-level comparisons. This pattern suggests that encouraging models to verbalize intermediate reasoning does not necessarily improve judgment quality. A plausible explanation is that, for subjective scoring tasks, explicit reasoning may introduce additional unsupported inferences and weaken calibration rather than strengthen it.

Obs. 3: Audio provides uniquely useful information for *Verbal Fluency*. Among the eight FIS dimensions, *Verbal Fluency* shows the clearest benefit from audio-aware evaluation. This is most evident when comparing audio+video+text against video+text settings under direct prompting: for all three models that support both configurations, adding audio improves ICC on *Verbal Fluency*, with a mean gain of 0.136, and also reduces MSE by an average of 0.178. This finding is consistent with the construct itself, since verbal fluency depends not only on lexical content but also on pace, hesitation, smoothness, and other prosodic cues that cannot be fully recovered from transcript or visual information alone.

Obs. 4: Current MLLMs have yet to fully and effectively leverage multimodal information. Richer multimodal input does not lead to consistent gains across models, and the best-performing configuration varies substantially from one model to another. In our experiments, the strongest setting for different models ranges from text-only direct and video+text direct to audio+text direct and audio+video+text CoT. The main bottleneck is therefore not merely access to more modalities, but the ability to integrate multimodal evidence and convert them into stable and dimension-specific judgments.

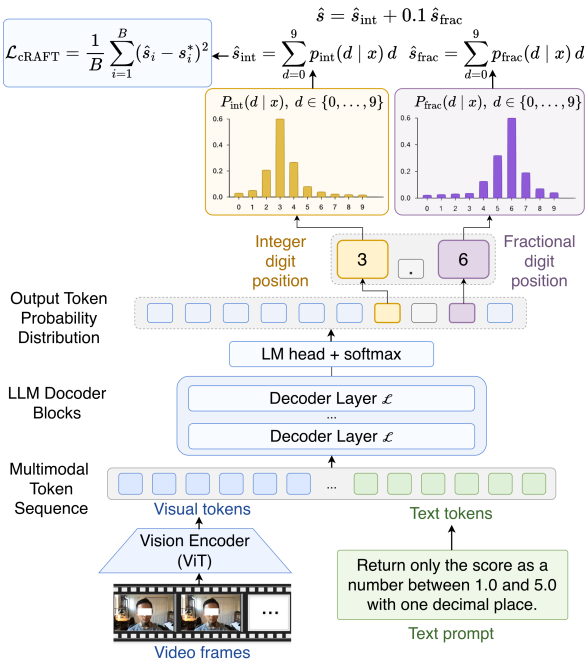


Figure 3: MIS-RAFT scoring pipeline for one-decimal score prediction.

5 MIS-RAFT

5.1 Motivation

In addition to the zero-shot and few-shot evaluations, we study whether task-specific adaptation can mitigate the misalignment between the autoregressive MLLM training objectives and the regression task here. In MIS-Bench, the target for each dimension is an averaged human score retained to one decimal place, which makes the problem fundamentally different from standard language generation. Under the common approach of token-level fine-tuning with cross-entropy loss, the decoder is rewarded for exact token matching and therefore treats score prediction as a task of string generation rather than numeric prediction. As a result, predictions such as "3.1" and "5.0" are both treated as equally incorrect relative to "3.0", even though they differ substantially in both clinical meaning and evaluation error. These considerations motivate a mitigation method that can directly optimize scalar score prediction instead of exact token matching. MIS-RAFT retains the standard autoregressive architecture and modifies only its training objective; it does not introduce an additional classification or

FIS Dimension	Raw	Fine-tuning	MIS-RAFT
VF	−3.8 [−12.5, 2.0]	**1.2** [−16.3, 18.9]	−0.7 [−16.0, 12.6]
H/PE	37.6 [24.0, 49.5]	21.7 [12.3, 31.2]	**50.7** [38.0, 62.3]
Per.	−2.0 [−6.6, 0.0]	1.8 [−15.7, 16.6]	**33.6** [20.2, 45.9]
Emo.	−0.0 [−0.0, 0.0]	−13.5 [−29.4, 4.1]	**17.0** [0.6, 31.1]
WAU	12.9 [−1.6, 27.3]	−12.2 [−22.9, 0.0]	**31.4** [14.9, 46.6]
Emp.	**43.6** [28.2, 55.9]	3.7 [−7.6, 15.3]	35.5 [19.1, 50.7]
ABC	15.2 [−5.3, 34.9]	7.6 [−16.6, 28.7]	**41.8** [26.8, 55.2]
ARRR	6.7 [−12.8, 25.6]	11.1 [−1.9, 24.2]	**39.2** [24.4, 52.4]
Mean Average	13.8	2.7	**31.1**

Table 4: ICC(3,1) by FIS dimension under raw inference, standard fine-tuning, and MIS-RAFT fine-tuning for the Qwen2.5-VL backbone. Values are reported in percentage points; brackets report 95% percentile bootstrap confidence intervals computed over videos using 10,000 bootstrap resamples. Higher is better, and the best ICC in each row is shown in bold.

regression head.

5.2 Method

To address the objective mismatch above, we propose **MIS-RAFT**, a regression-aware fine-tuning method for multimodal interpersonal skill scoring. MIS-RAFT is inspired by Regression-Aware Fine-Tuning (RAFT) proposed by Lukasik et al. (2025), which reformulates autoregressive numeric prediction as expected-value regression over the model's predictive distribution. Let x denote the input and s^* the gold score. Under standard token-level fine-tuning, the objective can be written as

$$\mathcal{L}_{\mathrm{CE}} = -\log p(\mathrm{str}(s^*) \mid x), \tag{6}$$

where $\mathrm{str}(s^*)$ is the string form of the target score. This objective encourages exact string matching, but does not encode the numerical distance between nearby predictions.

In the original RAFT formulation, a numeric prediction is constructed from the model distribution over a predefined candidate set $\mathcal{S}_{\mathrm{grid}}$:

$$\hat{s}_{\mathrm{RAFT}}(x) = \sum_{s \in \mathcal{S}_{\mathrm{grid}}} p(\mathrm{str}(s) \mid x)\, s, \tag{7}$$

and then minimizing the squared loss

$$\mathcal{L}_{\mathrm{RAFT}} = \big(s^* - \hat{s}_{\mathrm{RAFT}}(x)\big)^2. \tag{8}$$

Compared to cross-entropy, this objective is better aligned with scalar score prediction because it directly rewards numerically closer outputs.

However, this grid-based formulation is most natural when predictions are restricted to a finite, predefined set of target values. In MIS-Bench, scores are instead predicted at one-decimal precision, calling for a formulation that can directly accommodate more fine-grained values. We therefore extend RAFT from probability-weighted prediction over a single-token value grid to a compositional decimal representation, where integer and fractional digits jointly define the predicted score.

Specifically, as shown in Figure 3, MIS-RAFT decomposes each target score into an integer digit and a fractional digit. At inference, it computes probability-weighted expectations over the digit vocabulary $\{0, 1, \ldots, 9\}$:

$$\hat{s}_{\mathrm{int}} = \sum_{d=0}^{9} p_{\mathrm{int}}(d \mid x)\, d, \tag{9}$$

$$\hat{s}_{\mathrm{frac}} = \sum_{d=0}^{9} p_{\mathrm{frac}}(d \mid x)\, d. \tag{10}$$

Then combined into a continuous expected score prediction

$$\hat{s} = \hat{s}_{\mathrm{int}} + 0.1\, \hat{s}_{\mathrm{frac}}. \tag{11}$$

The final training objective minimizes the mean squared error between this expected score and the gold target:

$$\mathcal{L}_{\mathrm{cRAFT}} = \frac{1}{B} \sum_{i=1}^{B} (\hat{s}_i - s_i^*)^2. \tag{12}$$

5.3 Evaluation

Experimental Setup. We evaluate MIS-RAFT as a focused ablation study on top of MIS-Bench. Fol-

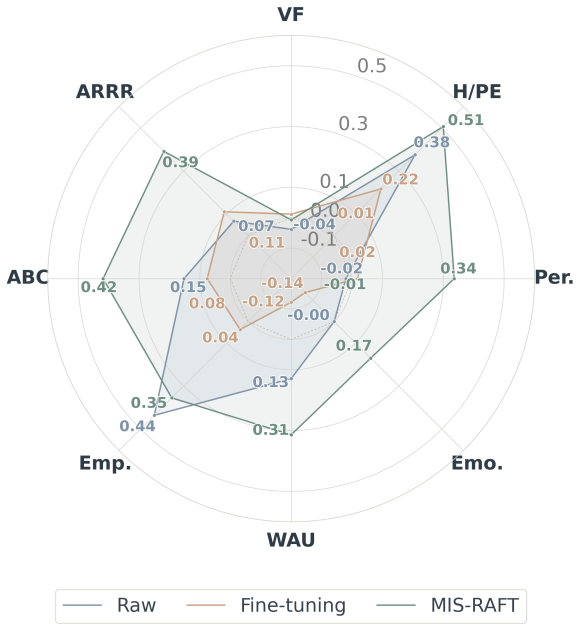


Figure 4: Radar chart of ICC scores across eight FIS dimensions.

lowing the benchmark results in the previous section, we use Qwen2.5-VL as the backbone model and apply LoRA-based parameter-efficient fine-tuning. Holding the backbone and PEFT framework fixed, we compare **raw inference** without task-specific tuning, **standard fine-tuning** with the conventional token-level cross-entropy objective, and **MIS-RAFT fine-tuning** with the regression-aware objective described above.

Result Analysis. Table 4 and Figure 4 show that MIS-RAFT achieves the strongest alignment with human ratings. Averaged across the eight dimensions, raw inference attains a mean ICC of 0.138, standard fine-tuning drops to 0.027, and MIS-RAFT improves to 0.311. MIS-RAFT outperforms substantially on most dimensions, especially in *Persuasiveness*, *Emotional Expression*, and *Alliance Bond Capacity*, indicating that the regression-aware objective is much better aligned with the nature of the task than standard token-level fine-tuning. Relative to the human single-rater ceiling in Table 2, MIS-RAFT approaches human agreement on selected dimensions, reaching 0.507 on *Hope and Positive Expectations* and 0.418 on *Alliance Bond Capacity*.

As shown in Figure 4, raw inference is uneven, with reasonable performance on a few dimensions but weak results on *Persuasiveness, Emotional Expression, and ARRR*. Standard fine-tuning introduces instability, including negative ICCs on some dimensions and a collapse on *Empathy*. In contrast, MIS-RAFT expands the profile more uniformly, which indicates broader and more balanced improvements.

6 Conclusion

This paper presents **MIS-Bench**, a benchmark for evaluating Multimodal Large Language Models on psychotherapy interpersonal skill assessment. Our experiments show that even the latest MLLMs, including both proprietary and open-source models, remain limited in their ability to align with expert human ratings on this expert-grounded task, with inconsistent gains from multimodal inputs, and even performance degradation under CoT prompting. We therefore introduce MIS-RAFT, a regression-aware mitigation for score prediction, and show that it improves alignment with expert ratings over standard fine-tuning. Future work could investigate a broader range of post-training strategies for this setting, including instruction tuning, preference-based alignment, and reinforcement learning across open-source MLLMs.

Limitations

Our study has several limitations. First, the sample size of the current benchmark is still modest relative to the diversity and complexity of real-world psychotherapy interactions. Second, models receive only the therapist response in the selected modality configuration, whereas human raters also view the standardized patient stimulus clip, although the rubric descriptions provide partial scenario context to compensate this model-rater information asymmetry. Third, while we study a focused regression-aware adaptation strategy through RAFT, we do not systematically investigate a broader range of post-training methods or alternative scoring formulations across open-source models, which would be a valuable direction for future work.

Ethical Considerations

This study evaluates MLLMs on psychotherapy interpersonal skill assessment in order to better understand their capabilities, limitations, and risks. All experiments are conducted in a benchmark setting using publicly available models or APIs, with no deployment in real clinical practice. Current models may still fail to align with expert judgment in systematic ways, and therefore not supposed to

be used for therapist supervision, training, hiring, or patient-facing evaluation, given their potential impact on the individuals being assessed and on downstream professional outcomes. MIS-Bench is intended to expose these limitations and support research on more responsible model-based evaluation.

Before data collection, all participants were informed of the study procedures, including the recording process and the intended research use of the collected data. Written informed consent was obtained from all participants for conversation recording and for the use of the resulting data in scientific publications, including conference and journal papers. The data were collected under a human subjects research protocol approved by the Institutional Review Board (IRB) of Yeshiva University, under Human Subject Research Protocol No. 1293523.

Acknowledgments

This work was supported by the Dean's Undergraduate Research Fund (DURF) and Center for Data Science at NYU Shanghai. Computational resources were provided by NYU Shanghai High Performance Computing and NYU Torch.

We are grateful to Jordan Bate of Adelphi University and Katie Aafjes-van Doorn for their contributions to the collection of FIS dataset. We also thank Bonian Jia and Kaifang Mao for helpful discussions and feedback.

References

Katie Aafjes-van Doorn, Marcelo Cicconet, Jordan Bate, Jeffrey F. Cohn, and Marc Aafjes. 2025. Development of an artificial intelligence-based measure of therapists' skills: A multimodal proof of concept. *Psychotherapy*, 62(3):301–314.

Katie Aafjes-van Doorn and Tao Lin. 2024. Will the facilitative interpersonal skills task reach its potential? *Clinical Psychology: Science and Practice*, 31(2):224–226.

J. J. Allen, A. Parker, and B. M. Ogles. 2024. A review of the facilitative interpersonal skills performance task and rating method. *Clinical Psychology: Science and Practice*, 31(2):211–223.

Timothy Anderson, M. E. J. Crowley, Lina Himawan, J. K. Holmberg, and B. D. Uhlin. 2016a. Therapist facilitative interpersonal skills and training status: A randomized clinical trial on alliance and outcome. *Psychotherapy Research*, 26(5):511–529.

Timothy Anderson, Joshua D. Finkelstein, and Sarah A. Horvath. 2020. The facilitative interpersonal skills method: Difficult psychotherapy moments and appropriate therapist responsiveness. *Counselling and Psychotherapy Research*, 20(3):463–469.

Timothy Anderson, Andrew S. McClintock, Lina Himawan, Xiaoxia Song, and Candace L. Patterson. 2016b. A prospective study of therapist facilitative interpersonal skills as a predictor of treatment outcome. *Journal of Consulting and Clinical Psychology*, 84(1):57–66.

Timothy Anderson, Benjamin M. Ogles, Candace L. Patterson, Michael J. Lambert, and David A. Vermeersch. 2009. Therapist effects: Facilitative interpersonal skills as a predictor of therapist success. *Journal of Clinical Psychology*, 65(7):755–768.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *Preprint*, arXiv:2402.04788.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Simon B. Goldberg, Michael Tanana, Zac E. Imel, David C. Atkins, Clara E. Hill, and Timothy Anderson. 2021. Can a computer detect interpersonal skills? using machine learning to scale up the facilitative interpersonal skills task. *Psychotherapy Research*, 31(3):281–288.

Google. 2026. Gemini api models: Gemini 3.1 pro preview. `https://ai.google.dev/gemini-api/docs/models`. Accessed: 2026-04-22.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on llm-as-a-judge. *The Innovation*, 7:101253.

Jared David Tadeo Guerrero-Sosa, Francisco P. Romero, Víctor Hugo Menéndez-Domínguez, Jesús Serrano-Guerrero, Andres Montoro-Montarroso, and José Angel Olivas. 2025. A multimodal framework for explainable evaluation of soft skills in educational environments. *CoRR*, abs/2505.01794.

He Hu, Lianzhong You, Hongbo Xu, Qianning Wang, Fei Richard Yu, Fei Ma, Zebang Cheng, Zheng Lian,

Yucheng Zhou, and Laizhong Cui. 2025. Emobench-m: Benchmarking emotional intelligence for multimodal large language models. *arXiv preprint arXiv:2502.04424*.

Jongwoo Ko, Sungnyun Kim, Sungwoo Cho, and Se-Young Yun. 2025. Flex-judge: Text-only reasoning unleashes zero-shot multimodal evaluators. *arXiv preprint arXiv:2505.18601*.

Woosuk Kwon, Zijia Yin, Beomjin Han, Jiale Chen, Rui Lo, Tianqi Wu, Huan Yu, Guangyu Shi, Sijie Sheng, Fan Zhang, Lianmin Zhou, Yitao Fei, Xinfeng Tan, Cheng Xie, Haoran Li, Sicun Liu, Han Yao, Yang You, Hao Zhang, and 1 others. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626. Association for Computing Machinery.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Michal Lukasik, Zhao Meng, Harikrishna Narasimhan, Yin-Wen Chang, Aditya Krishna Menon, Felix Yu, and Sanjiv Kumar. 2025. Better autoregressive regression with llms via regression-aware fine-tuning. In *International Conference on Learning Representations*.

OpenAI. 2026. Introducing gpt-5.4. `https://openai.com/index/introducing-gpt-5-4/`. Accessed: 2026-04-22.

OpenBMB. 2026. Minicpm-o 4.5. `https://huggingface.co/openbmb/MiniCPM-o-4_5`. Hugging Face model card. Accessed: 2026-04-25.

Sowmya Rasipuram and Dinesh Babu Jayagopi. 2020. Automatic multimodal assessment of soft skills in social interactions: a review. *Multimedia Tools and Applications*, 79(19–20):13037–13060.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges. *arXiv preprint arXiv:2406.12624*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

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

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

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, and 15 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

James M. Zech, Robert Steele, Victoria K. Foley, and Thomas D. Hull. 2022. Automatic rating of therapist facilitative interpersonal skills in text: A natural language processing application. *Frontiers in Digital Health*, 4:917918.

Fan Zhang, Zebang Cheng, Chong Deng, Haoxuan Li, Zheng Lian, Qian Chen, Huadai Liu, Wen Wang, Yi-Fan Zhang, Renrui Zhang, Ziyu Guo, Zhihong Zhu, Hao Wu, Haixin Wang, Yefeng Zheng, Xiaojiang Peng, Xian Wu, Kun Wang, Xiangang Li, and 2 others. 2025. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv preprint arXiv:2508.09210*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Implementation Details

A.1 Prompt Templates

The eight prompt templates used in the main benchmark are provided below, covering the four modality configurations and two prompting styles.

Prompt for T-direct

```
You are an expert evaluator of clinical communication skills.

### Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided transcript only.

Use only the transcript; do not assume visual or audio cues.

### Dimension:
{dimension_name}

### Response to evaluate:
{transcript}

### Score Rubrics:
{score_rubric}

### Instructions:
Return only the score as a number between 1.0 and 5.0 with one decimal place.
```

Prompt for T-CoT

```
You are an expert evaluator of clinical communication skills.

### Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided transcript only.

Use only the transcript; do not assume visual or audio cues.

### Dimension:
{dimension_name}

### Response to evaluate:
{transcript}

### Score Rubrics:
{score_rubric}

### Instructions:
1. Reasoning Process (<think>): Provide a detailed analysis inside <think> tags.
You must:
- Use only transcript evidence.
- Do not assume any auditory or visual evidence.
- Compare the observed evidence strictly against the score descriptions provided above.
- Explain the logical steps that lead to your final score.
- You may assign any score from 1.0 to 5.0 with one decimal place.
2. Final Answer: Output exactly one line:
So the predicted score is X.X.
(1.0 <= X.X <= 5.0, one decimal place)

### Response Template:
<think>
[Detailed reasoning process based on transcript evidence only]
</think>
So the predicted score is X.X.
```

Prompt for AT-direct

```
You are an expert evaluator of clinical communication skills.

<audio>

### Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided audio recording and transcript.

Analyze auditory cues (tone, pacing, speech style).

### Dimension:
{dimension_name}

### Response to evaluate:
{transcript}

### Score Rubrics:
{score_rubric}

### Instructions:
Return only the score as a number between 1.0 and 5.0 with one decimal place.
```

Prompt for AT-CoT

```
You are an expert evaluator of clinical communication skills.

<audio>

### Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided audio recording and transcript.

Analyze auditory cues (tone, pacing, speech style).

### Dimension:
{dimension_name}

### Response to evaluate:
{transcript}

### Score Rubrics:
{score_rubric}

### Instructions:
1. Reasoning Process (<think>): Provide a detailed analysis inside <think> tags.
You must:
- Use only transcript and auditory evidence (tone, pacing, speech style).
- Do not assume visual or non-verbal video evidence.
- Compare the observed evidence strictly against the score descriptions provided above.
- Explain the logical steps that lead to your final score.
- You may assign any score from 1.0 to 5.0 with one decimal place.
2. Final Answer: Output exactly one line:
So the predicted score is X.X.
(1.0 <= X.X <= 5.0, one decimal place)

### Response Template:
<think>
[Detailed reasoning process based on transcript and audio evidence]
</think>
So the predicted score is X.X.
```

Prompt for VT-direct

You are an expert evaluator of clinical communication skills.

<video>

Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided video recording and transcript.

Analyze visual cues (facial expressions, body language).

Dimension:
{dimension_name}

Response to evaluate:
{transcript}

Score Rubrics:
{score_rubric}

Instructions:
Return only the score as a number between 1.0 and 5.0 with one decimal place.

Prompt for VT-CoT

You are an expert evaluator of clinical communication skills.

<video>

Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided video recording and transcript.

Analyze visual cues (facial expressions, body language).

Dimension:
{dimension_name}

Response to evaluate:
{transcript}

Score Rubrics:
{score_rubric}

Instructions:
1. Reasoning Process (<think>): Provide a detailed analysis inside <think> tags.
You must:
- Use only transcript and visual evidence (facial expressions, body language).
- Do not assume audio prosody, tone, or pacing evidence.
- Compare the observed evidence strictly against the score descriptions provided above.
- Explain the logical steps that lead to your final score.
- You may assign any score from 1.0 to 5.0 with one decimal place.
2. Final Answer: Output exactly one line:
So the predicted score is X.X.
(1.0 <= X.X <= 5.0, one decimal place)

Response Template:
<think>
[Detailed reasoning process based on transcript and visual evidence]
</think>
So the predicted score is X.X.

Prompt for AVT-direct

You are an expert evaluator of clinical communication skills.

<video> <audio>

Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided video recording and audio.

Analyze both visual cues (facial expressions, body language) and auditory cues (tone, pacing, speech style).

Dimension:
{dimension_name}

Response to evaluate:
{transcript}

Score Rubrics:
{score_rubric}

Instructions:
Return only the score as a number between 1.0 and 5.0 with one decimal place.

Prompt for AVT-CoT

You are an expert evaluator of clinical communication skills.

<video> <audio>

Task Description:
You are an expert evaluator of clinical communication skills. Assess the therapist's Facilitative Interpersonal Skills (FIS) for ONE dimension only based on the provided video recording and audio.

Analyze both visual cues (facial expressions, body language) and auditory cues (tone, pacing, speech style).

Dimension:
{dimension_name}

Response to evaluate:
{transcript}

Score Rubrics:
{score_rubric}

Instructions:
1. Reasoning Process (<think>): Provide a detailed analysis inside <think> tags.
You must:
- Use transcript, auditory evidence, and visual evidence together.
- If modalities conflict, explain which evidence is more reliable and why.
- Compare the observed evidence strictly against the score descriptions provided above.
- Explain the logical steps that lead to your final score.
- You may assign any score from 1.0 to 5.0 with one decimal place.
2. Final Answer: Output exactly one line:
So the predicted score is X.X.
(1.0 <= X.X <= 5.0, one decimal place)

Response Template:
<think>
[Detailed reasoning process based on transcript, audio, and visual evidence]
</think>
So the predicted score is X.X.

A.2 Score Rubrics

Table 5: Full scoring rubrics for the eight FIS dimensions.

Verbal Fluency
Criterion. Verbal Fluency. This item is a rating of the extent to which the participant is verbally capable and at-ease in communicating. The response is delivered in a relaxed manner and without significant signs of anxiety (e.g., broken speech, extended and awkward pauses, and clarity in communication). However, the content of what is said is not rated, but rather how it is spoken.
Score 1. The participant has great difficulty verbalizing a response (e.g., obviously anxious, sounds shaky or timid), reflecting a clear avoidance or anxiety. The participant may lack confidence and is clearly uncertain or even difficult to follow.
Score 2. Fluency is disrupted by the participant's anxiety and avoidance of verbal expression. The respondent may be obviously anxious and struggling to formulate a response. At times, the communication may be choppy, even halting. [Note: In some rare instances a response could represent an avoidance of the interpersonal situation through anxious rambling. It would need to be clear that the participant's ramblings are the result of anxiety over communicating with another].
Score 3. A moderate level of verbal fluency indicates that the participant's response is conversational and mostly easy to follow.
Score 4. The response is fluent, and there is little that is difficult to follow.
Score 5. The participant is at great ease and communicates ideas with no anxiety, reflecting a desire to "approach" the other. The verbal quality of the response may have a "melodic," rhythmical quality and is easy to follow; the response is fluent.

Hope & Positive Expectations
Criterion. Hope & Positive Expectations. This item rates expressions of hope, optimism, and positive expectations for change. Staats (1989, 2001) defines hope as the interaction between wishes and expectations. The interpersonal skills needed for hope involve facilitating a) personal agency and b) building the pathways needed for attaining desired goals and expectations (Steed, 2002). Hope is related to persuasiveness and collaboration in the sense that hope and positive expectations are often built through offering a rationale, friendliness, and enthusiasm. As defined here, hope focuses more on building client agency for actions that will facilitate meeting the client's goals whereas persuasion is based more on a plausible explanation (which may or may not include hope).
Score 1. The participant's response is hopeless or is even pessimistic. For example, the participant may address only issues or concerns beyond the control of the other or subtly suggests that the other cannot change or improve his/her problems.
Score 2. The participant responds with some hopelessness, including subtle expressions of feeling unable to help the client.
Score 3. The response is ordinary OR the optimism of the response is not discernible. There may be some hopefulness expressed, but with little confidence or reason for being hopeful.
Score 4. A general sense of optimism about the client's situation is detected. Specifically, the participant's response is directed toward building the client's agency OR facilitating the building of pathways to meet the client's goals.
Score 5. The participant's response expresses clear hope about the client's future and/or positive expectations about therapeutic work. In addition, for a response to be coded as a "5" there needs to be an allusion to building the client's agency as well as how the client might participate or do something that will help move toward his/her desired goals (i.e., pathways).

Persuasiveness
Criterion. Persuasiveness. Persuasiveness is the capacity to induce the other to accept a view that may be different from his or her own view. It involves that ability to convey a clear, organized understanding about the meaning of the other's source of distress. Persuasiveness implies an ability to communicate what Jerome Frank called a "believable myth." This capacity implies that the persuasive therapist must be convincing in communicating this belief-system. Rating Notes: For rating purposes, the response does not necessarily need to convey an entire world view, but a point of view that is implied to be at least slightly different from the client in the video clip. High ratings require that the participant provide a clear belief in a point of view or rationale. It is necessary that the rationale be relevant to the other's problems and at least somewhat novel to the other's experience. For this item, the rater should disregard personal beliefs about the validity of the participant's rationale, but instead rate the extent to which the participant might persuade another (i.e., ability to "sell" their rationale).
Score 1. The participant's response is unorganized, incoherent, and difficult to follow. The participant may also not know what to say.
Score 2. The participant is unpersuasive. Unpersuasive responses may be characterized by either
Score 3. The participant's response conveys little sense of persuasiveness.
Score 4. The participant speaks persuasively. The rationale may be more implicit and it is even possible that the rationale, though present, may be unclear, superficial, or marginally relevant to the other's problems.
Score 5. The participant is highly persuasive. Persuasive persons may speak with great confidence, certainty, and authority. Advice may or may not be given, but the participant must offer some rationale or re-framing of the other's experience.

Emotional Expression
Criterion. Emotional Expression. This item rates the energy and emotion in the participant's response. This item rates the extent to which the participant's response is delivered with effective expressions of emotion.
Score 1. The participant speaks with little or no affect and may be dull or boring (e.g., speaking in a near monotone voice and without intensity).
Score 2. The participant may display some sense of interest or curiosity, but the response is not emotionally engaging. Prosody is somewhat less than typical to casual conversation.
Score 3. The participant has prosody, but it is the amount of emotion that one might find in ordinary conversation.
Score 4. The participant is emotionally expressive at a moderate level. There is more emotion than found in ordinary speech, but it is not as focused in its delivery as the maximum rating of 5.
Score 5. There is affect and prosody in the participant's voice. The response is delivered in a highly emotional and engaging manner. The primary criterion is that the vocal expression conveys emotion. There may be a more focused delivery of emotional intonations to emphasize meanings that influence other processes (e.g. persuasion). The participant may even be somewhat provocative or challenging in delivering an emotion-based response toward the client in the video clip. However, a "5" should not be rated if the affect is primarily demeaning or hostile toward the other (in which case a "3" would be the maximum rating possible).

Warmth, Acceptance, & Understanding
Criterion. Warmth, Acceptance, & Understanding. This item is a rating of the ability of the participant to care for and accept the other. Therapist behaviors/attitudes that might indicate an absence of acceptance and understanding include: a judgmental attitude, condescension, rudeness, disapproval, guilt-induction, exasperation, or annoyance. Often it will be necessary to avoid rating what the participant is doing (e.g., giving advice), but rate how it is being done. Note that accepting does not necessarily mean approval, but rather a caring attitude and determination to help the other.
Score 1. The participant has an obvious lack of respect, acceptance, or warmth for the other (e.g., clearly pejorative comments, judgmental attitude, condescension, disapproval, guilt induction, blaming the other).
Score 2. The participant conveys a subtle lack of respect, acceptance, or concern of the other (e.g., sarcasm, exasperation, annoyance).
Score 3. There is an "ordinary" level of courtesy and warmth in the response OR the participant's opinion of the other may not be clearly discernable from the response.
Score 4. The participant's response is genuinely nonjudgmental and gently explores the other's thoughts, feelings, alternatives for dealing with future situations, etc. The participant appears concerned for and respectful of the client.
Score 5. The participant expresses clear and obvious warmth, concern and acceptance. The participant may, for example, make a compassionate attempt to relate to the other's experience.

Empathy
Criterion. Empathy. The capacity to respond with an expressed understanding of the subjective experience of the client. The response must also convey an accurate understanding of the thoughts and emotions expressed in the video clip. Therefore, it is especially important that the rater have an accurate understanding of the client's experiences in the video clips.
Score 1. Participant clearly distorts the other's experience. That is, the participant misidentifies a significant component of the other's complaints, beliefs, emotions, etc. Give a rating of 1 if the response indicates a clear disregard of the other's experience.
Score 2. Participant does not communicate an awareness or understanding of the other's experience, and/or there are minor distortions of the other's experience. Some aspects of the participant's response may be irrelevant to the other's concerns (when clearly not an attempt to change the other's focus).
Score 3. Participant is generally accurate about the other's experience but only perceives the more obvious aspects of the other's experience or concerns.
Score 4. Participant comments accurately on the other's experience but not to the extent required to receive a "5" rating. The distinction between the 4 and 5 ratings are matters of intensity.
Score 5. Participant alludes to the client's experience so that it is clear that he/she has not only listened, but obtained an exceptional comprehension of what the other is experiencing. In order to receive a "5" the participant must infer something about the other's experience that is not explicitly stated by the other but is clearly identifiable in the client's nonverbal expression.

Alliance Bond Capacity
Criterion. Alliance Bond Capacity. This item rates the participant's capacity to provide a collaborative environment, one in which there is recognition of the need to work with the client jointly on problems.
Score 1. The participant actively undermines a mutual collaboration. The participant may respond in a way that is over-involved or reactive (e.g., moralistic lecturing, "preaching" to the other, assuming all responsibility). The rupture may also involve withdrawal or under-involvement in the participant's response (e.g., putting all the responsibility for change on the other).

Score 2. The participant may slightly undermine the building of a collaborative atmosphere, although it may be unintentional or superficial.
Score 3. The participant neither undermines nor attempts to enhance a collaborative effort.
Score 4. Some effort to collaborate is made but not as strong as a "5" (e.g., subtle invitations to engage in working with the client).
Score 5. Specific actions on the part of the participant help create a collaborative atmosphere. There should be a sense that the participant is attempting to work with the other to create a "we-ness" that is implied in the participant's behavior (e.g., participant checks with the other by asking questions about the "fit" of interpretations, conclusions, goals, etc.).

Alliance Rupture-Repair Responsiveness
Criterion. Alliance Rupture-Repair Responsiveness. Background: Each client in video clips is expressing an interpersonal issue that involves the patient-therapist relationship. Each video clip places the participant in the middle of alliance rupture episodes. Further, these rupture episodes take place at different locations within the interpersonal circle, which requires interpersonal flexibility for the therapist. The interpersonal problem with the client-therapist relationship threatens the development of the alliance. This item rates the extent to which the therapist appears responsive to the interpersonal issue. In some cases, the problem is clearly stated as when Suzie angrily berates the therapist for being ineffective. In other cases, the problem is more implicit such as when Lauren idealizes the therapist to the extent of leaving herself overly vulnerable to disappointment.
Score 1. Participant reacts to the interpersonal tension in a way that is nonproductive or in a way that likely exacerbates the rupture (e.g., responding negatively to a hostile client; responding to a controlling client with counter-control). Low scores also may be given when the participant fails to respond to the interpersonal issue involved in a way that indicates that the participant is avoiding the interpersonal issue or the relationship altogether.
Score 2. Participant addresses an issue to discuss that it is tangentially related to the interpersonal issue presented, but directs the discussion away from the present relationship situation.
Score 3. There may be more casual recognition of the interpersonal situation, but the response does not draw for further discussion of the issue or the relationship.
Score 4. Participant recognizes the other's interpersonal issue, and may discuss this further in more general terms (or discuss some secondary element of the other's issue or the relationship).
Score 5. Participant makes attempts to repair the interpersonal issue by engaging the client in a direct discussion of the immediate moment-to-moment interaction. This may include how specific relational messages are being expressed by the client in the video clip. Optimal responses will include attempts to not only allude to the interpersonal tension, but make some attempt to repair that interpersonal issue.

B Annotation Protocol and Human-Rater Reliability

Each video was independently rated by trained expert raters following the original FIS rating manual. Raters were trained by licensed clinicians who had themselves been trained by Tim Anderson, the developer of the FIS manual. Training continued until each rater achieved ICC greater than 0.8 with expert ratings on six consecutive clips. Raters also attended weekly calibration meetings to discuss ratings and prevent rater drift.

C Dataset Statistics

We report descriptive statistics of the FIS benchmark labels (gold scores), including inter-dimension correlation, per-dimension summary statistics, and gold-score distributions. The dataset contains 996 unique videos and 7,968 gold-score annotations (8 dimensions per video). The average video duration over all 996 videos is 62.42 seconds.

Inter-dimension Correlation. Figure 5 shows the Pearson correlation matrix across the eight FIS dimensions, computed from gold scores at the sample level.

Per-dimension Summary Statistics. Table 6 reports the mean, standard deviation, minimum, and maximum of gold scores for each FIS dimension.

Gold Score Distributions. Figure 6 shows the gold-score histograms for all eight FIS dimensions.

D Additional Experimental Results

D.1 MIS-RAFT Setup

In our MIS-RAFT experiments, we fine-tune the base model Qwen/Qwen2.5-VL-7B-Instruct with 4-bit NF4 quantization and PEFT LoRA on two A100 GPUs. For both MIS-RAFT and the conventional SFT training objective, the model is trained for 2 epochs with a learning rate of 2×10^{-5}, per-device training batch size of 2, maximum sequence length of 512, and bfloat16 computation. We use the `paged_adamw_32bit` optimizer. The dataset is split at the **video level** into train, validation, and test sets with an 80/10/10 ratio, ensuring that no video appears in more than one split. For video preprocessing, we set the total pixel budget to $20480 \times 28 \times 28$ and the minimum pixel budget to $16 \times 28 \times 28$.

Table 6: Per-dimension summary statistics of FIS gold scores.

Dimension	Mean	Std	Min	Max
VF	3.658	0.493	1.0	5.0
H/PE	3.420	0.394	2.0	4.3
Per.	3.613	0.454	1.3	5.0
Emo.	3.627	0.426	2.0	5.0
WAU	3.607	0.443	1.3	5.0
Emp.	3.534	0.414	2.2	4.7
ABC	3.564	0.478	1.3	5.0
ARRR	3.433	0.585	1.0	5.0

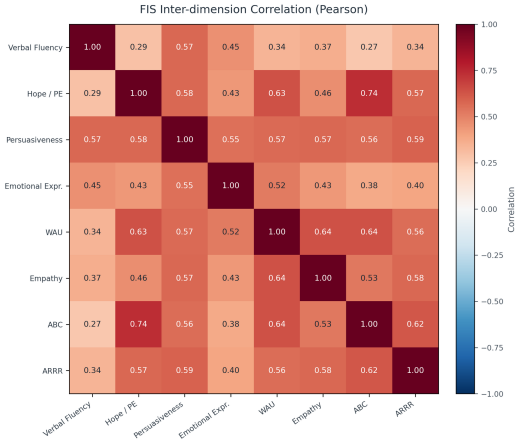


Figure 5: Inter-dimension Pearson correlation of FIS gold scores.

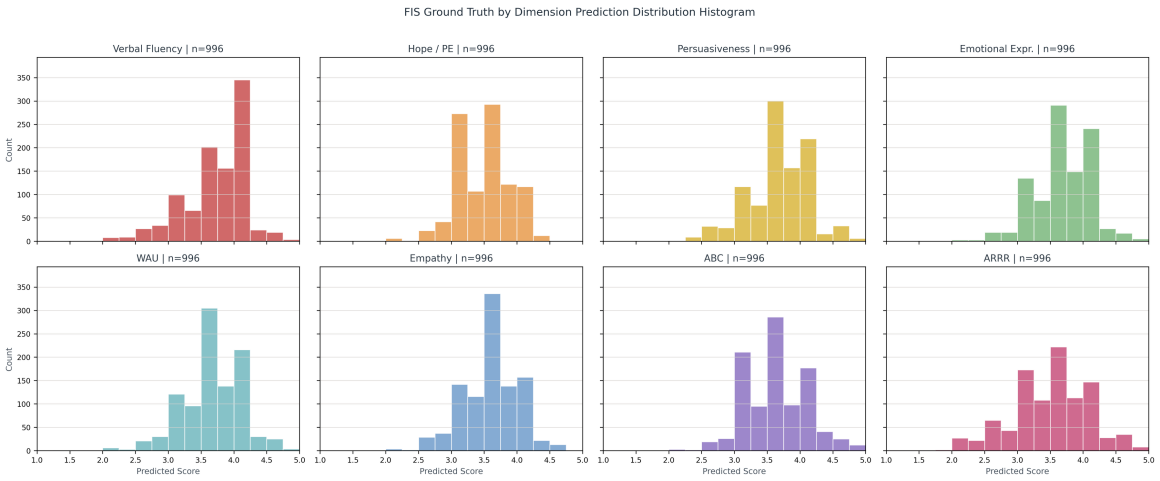


Figure 6: Gold-score histograms by dimension on the FIS dataset.

D.2 Additional Dimension-Level Heatmaps

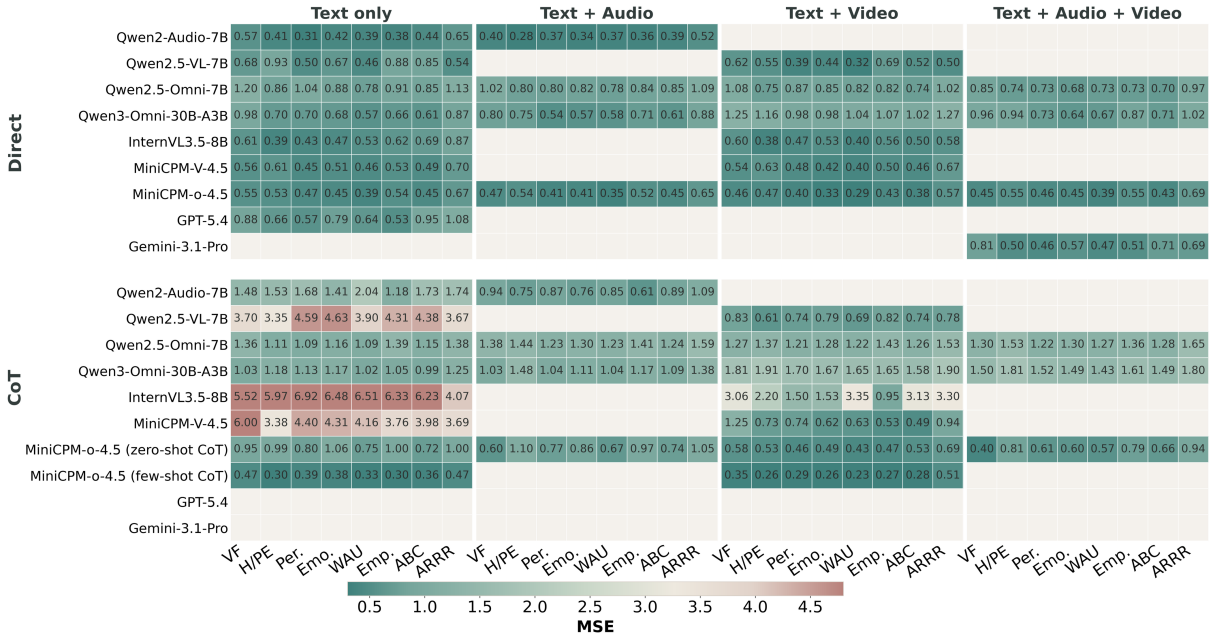


Figure 7: **Dimension-level MSE heatmaps on MIS-Bench.** Lower MSE indicates smaller prediction error with respect to human ratings.

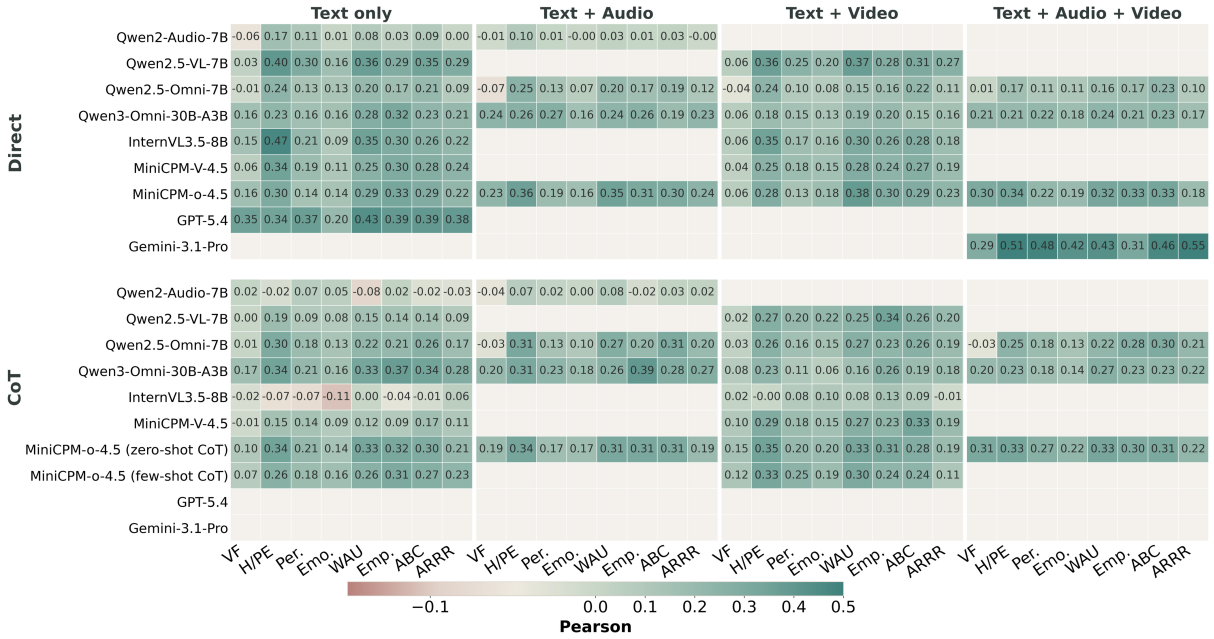


Figure 8: **Dimension-level Pearson heatmaps on MIS-Bench.**

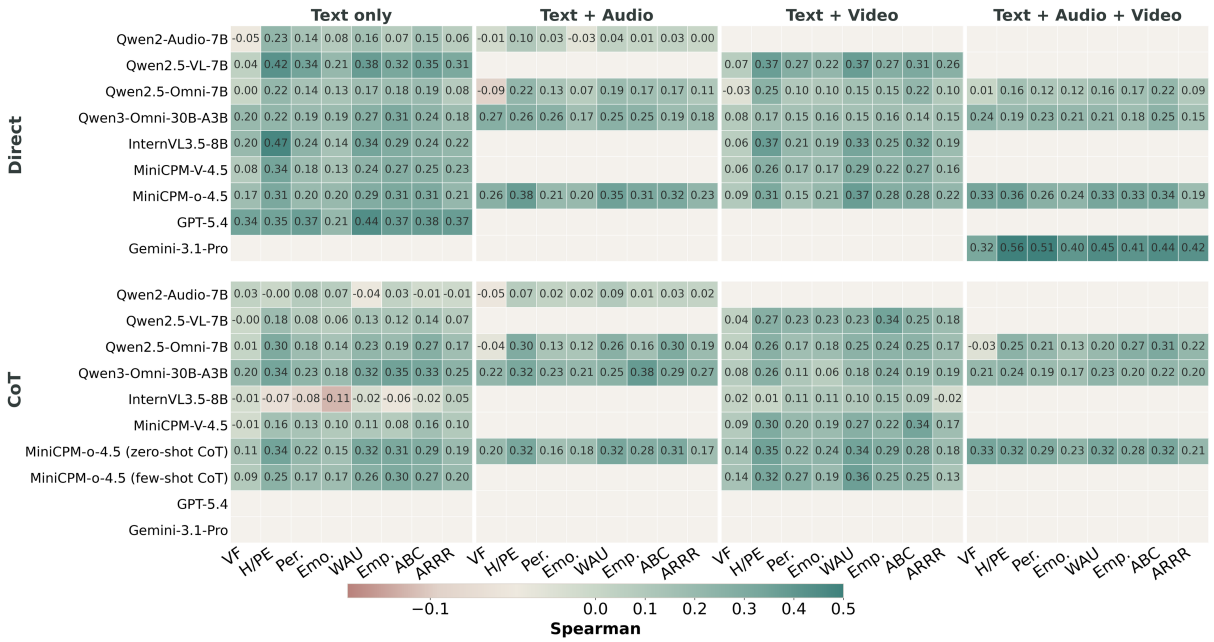


Figure 9: **Dimension-level Spearman heatmaps on MIS-Bench.**

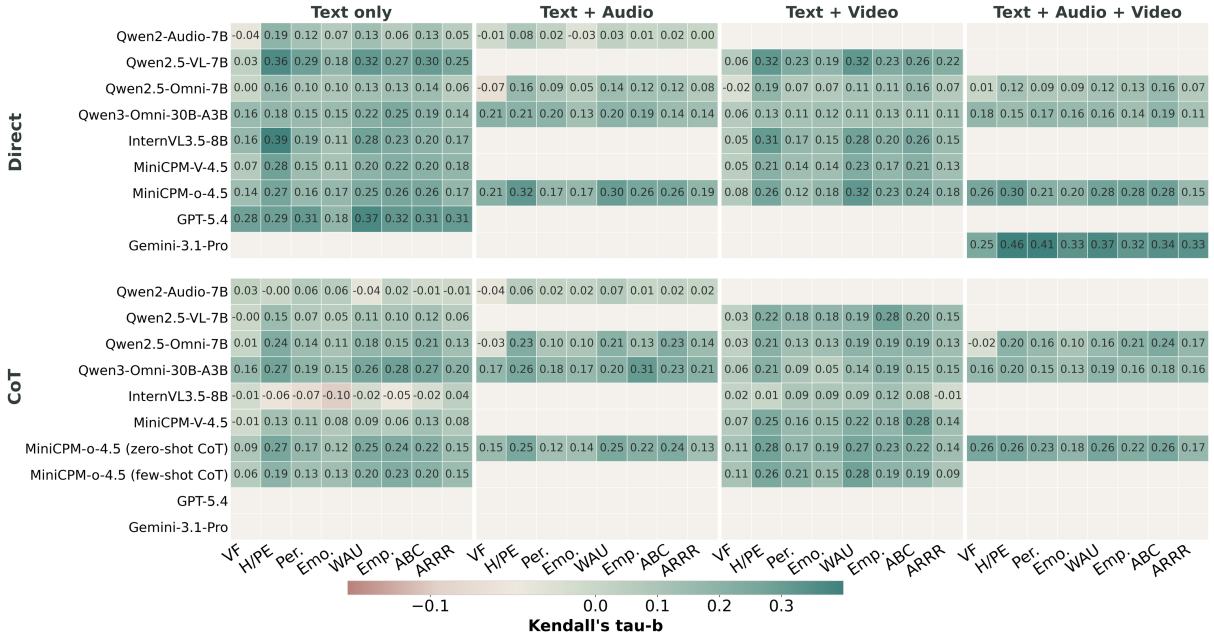


Figure 10: **Dimension-level Kendall's tau-b heatmaps on MIS-Bench.**

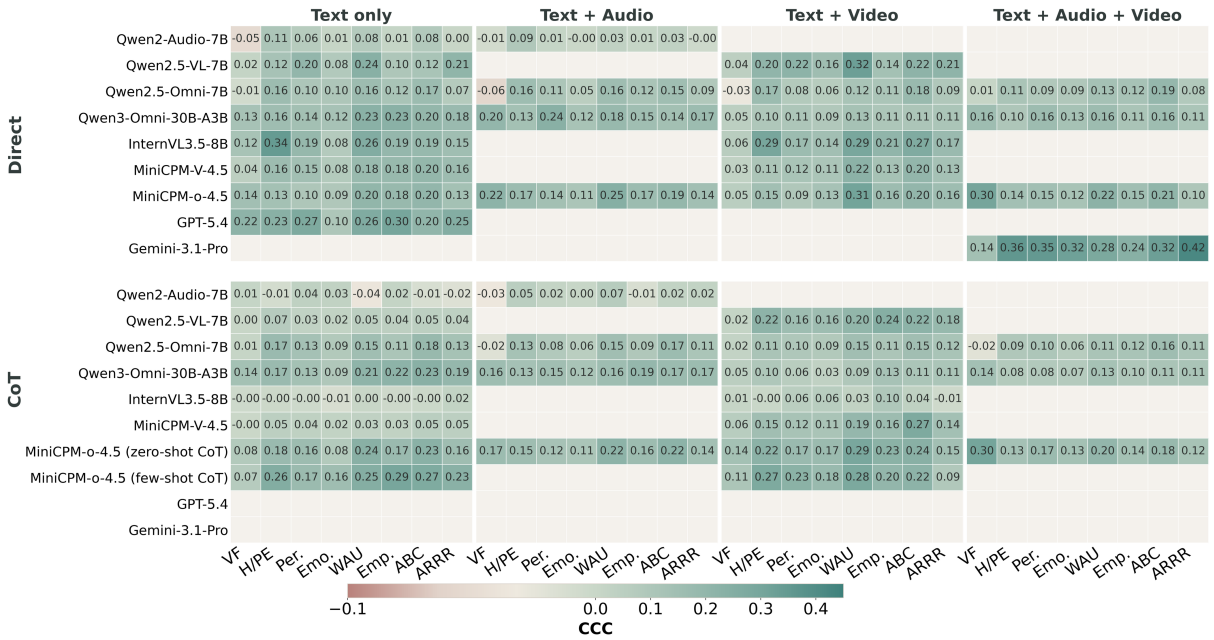


Figure 11: **Dimension-level CCC heatmaps on MIS-Bench.**

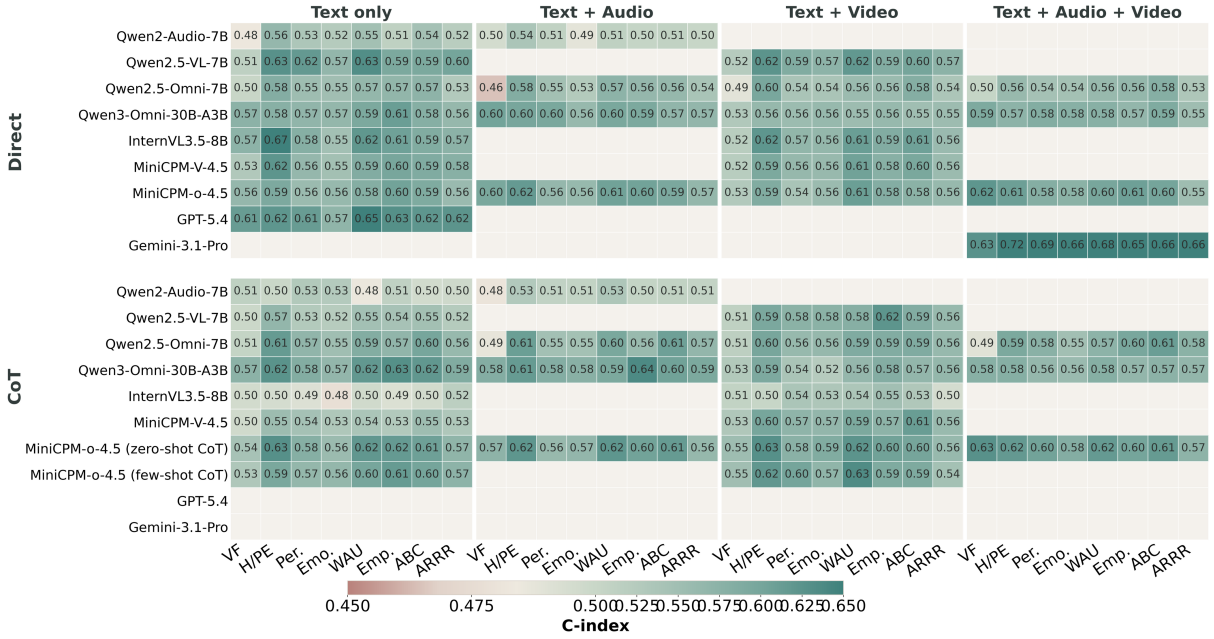
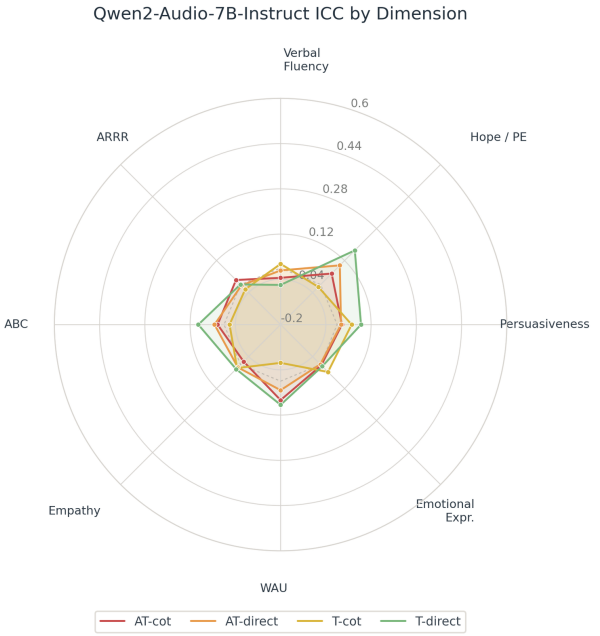


Figure 12: **Dimension-level C-index heatmaps on MIS-Bench.**

D.3 Model-wise Radar Plots

D.3.1 ICC Radar Plots



Qwen2-Audio-7B-Instruct



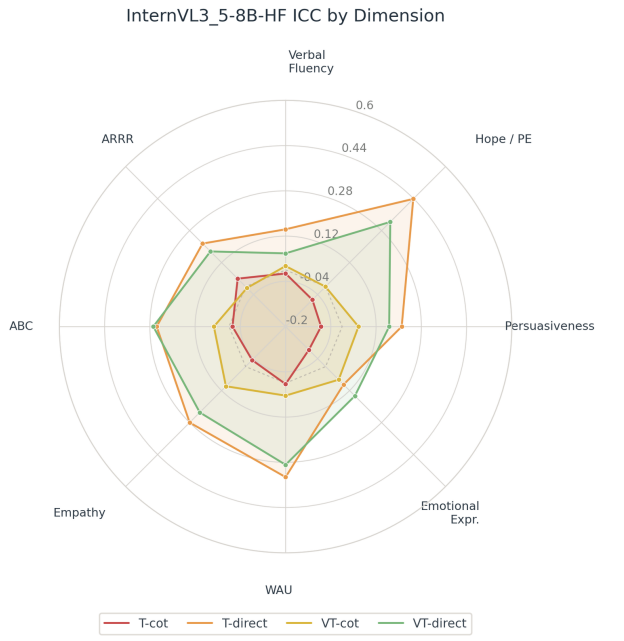
Qwen2.5-VL-7B-Instruct



Qwen2.5-Omni-7B



Qwen3-Omni-30B-A3B-Instruct



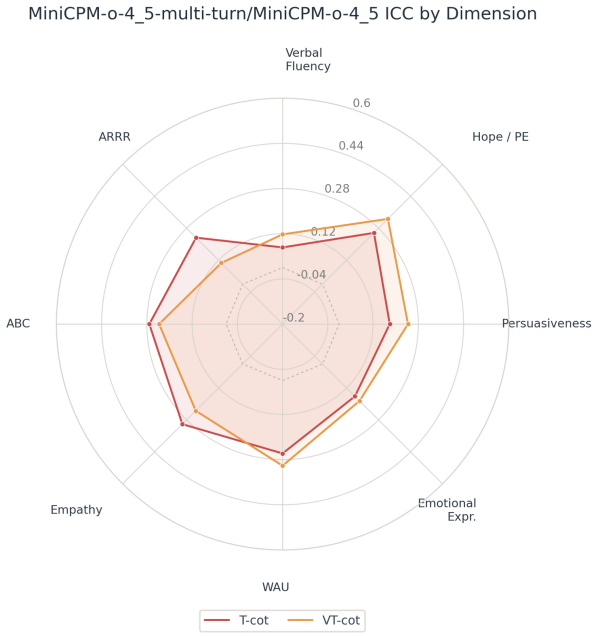
InternVL3.5-8B-HF



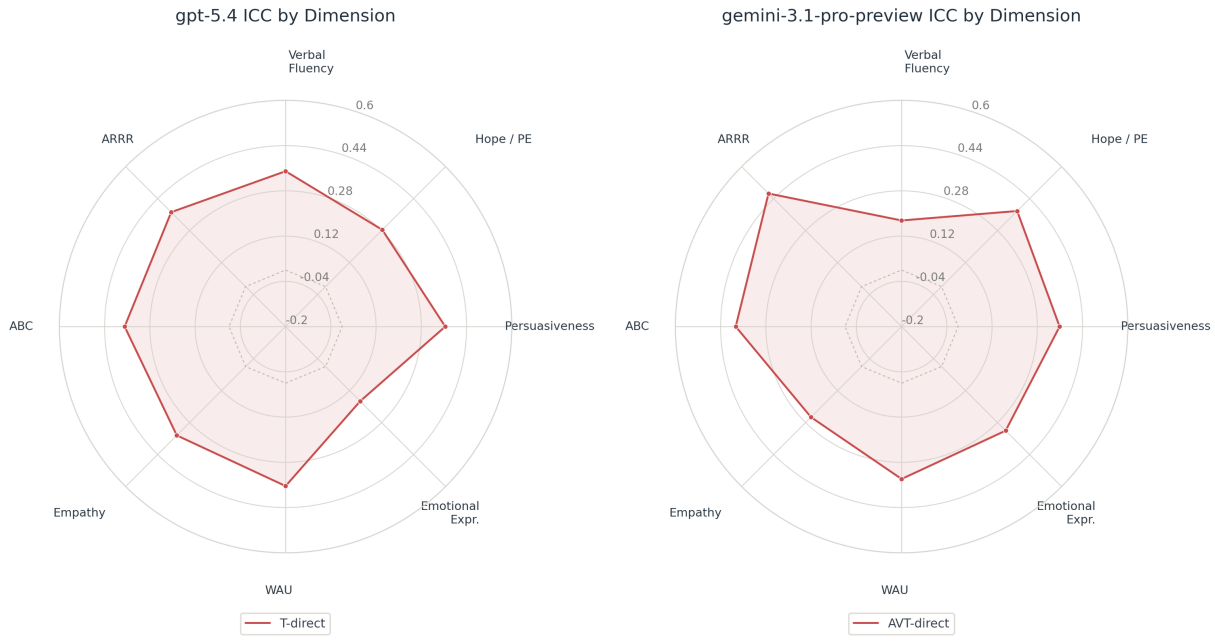
MiniCPM-V-4.5



MiniCPM-o-4.5



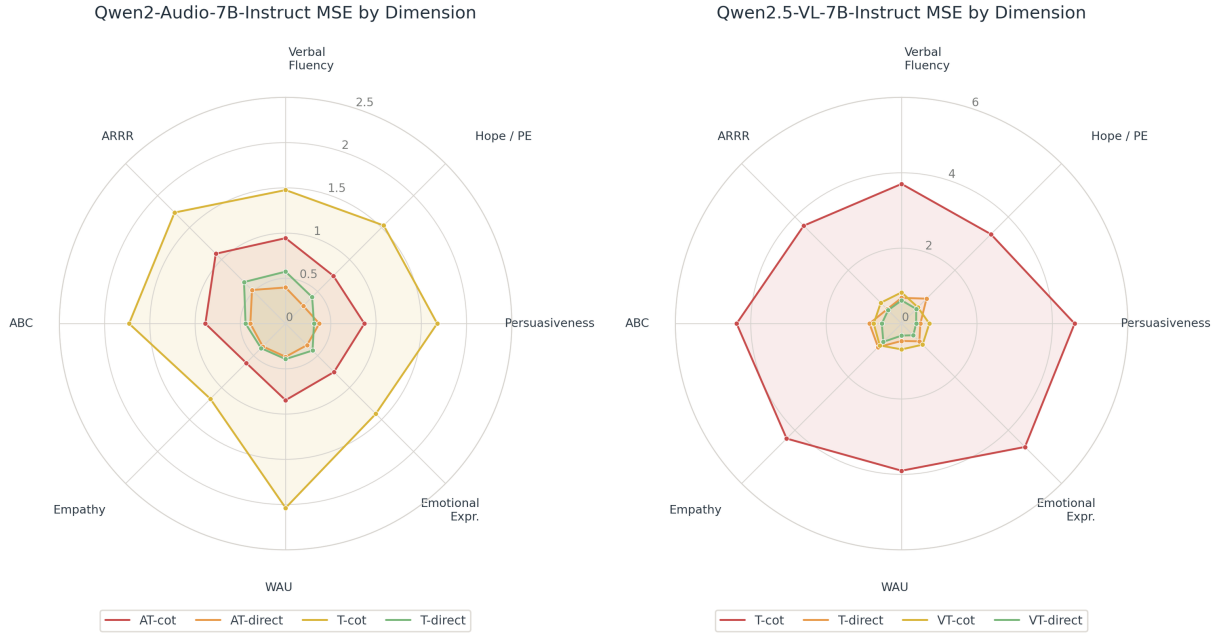
MiniCPM-o-4.5 multi-turn



GPT-5.4 Gemini-3.1-Pro-Preview

Figure 13: **Model-wise ICC radar plots.**

D.3.2 MSE Radar Plots



Qwen2-Audio-7B-Instruct Qwen2.5-VL-7B-Instruct



Qwen2.5-Omni-7B



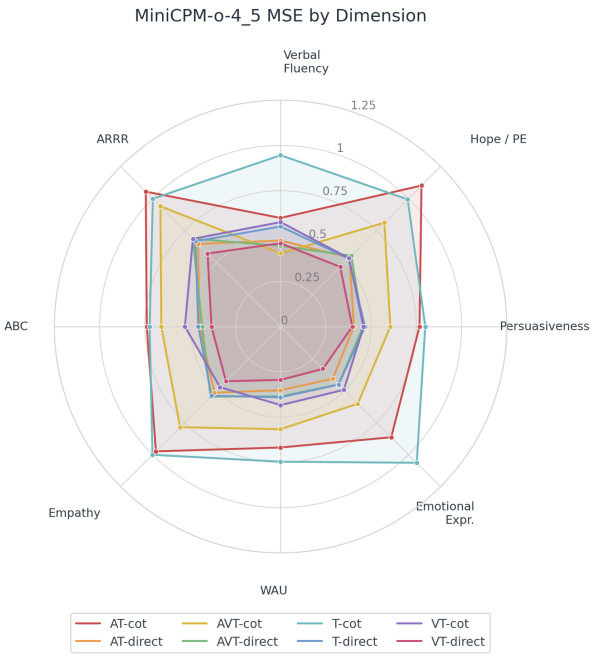
Qwen3-Omni-30B-A3B-Instruct



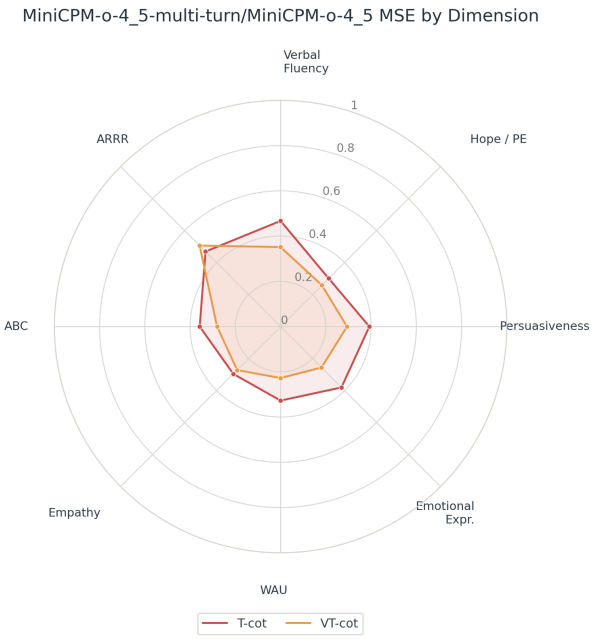
InternVL3.5-8B-HF



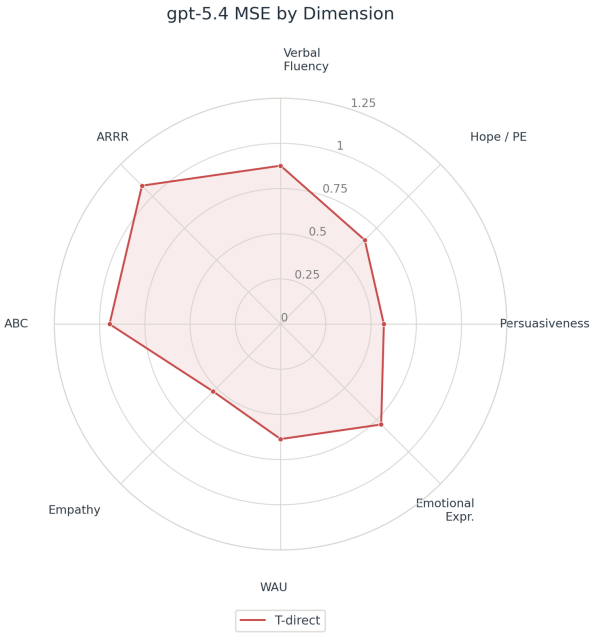
MiniCPM-V-4.5



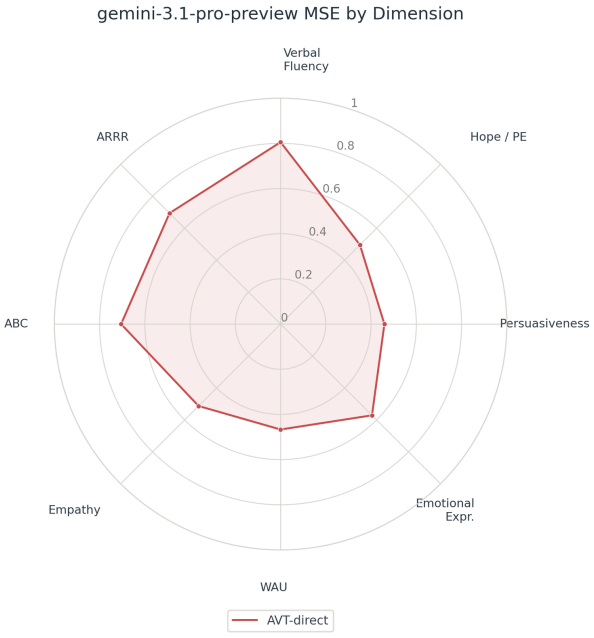
MiniCPM-o-4.5



MiniCPM-o-4.5 multi-turn



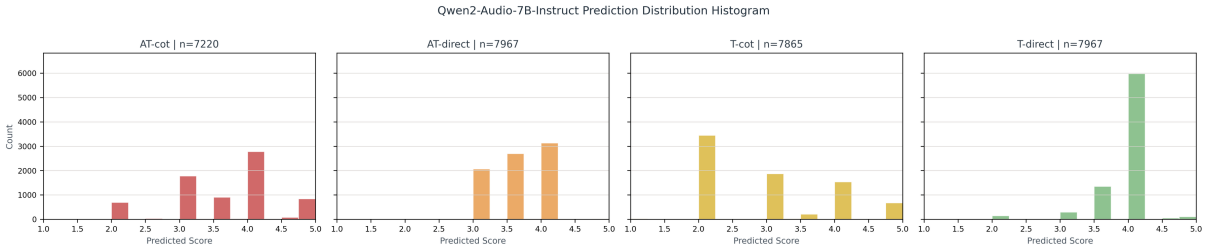
GPT-5.4



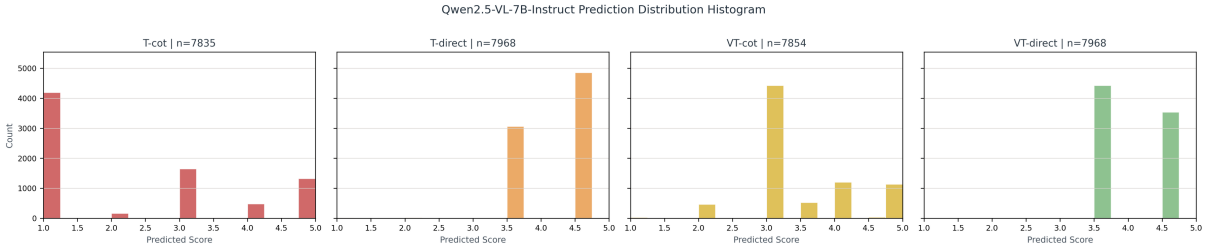
Gemini-3.1-Pro-Preview

Figure 14: **Model-wise MSE radar plots.**

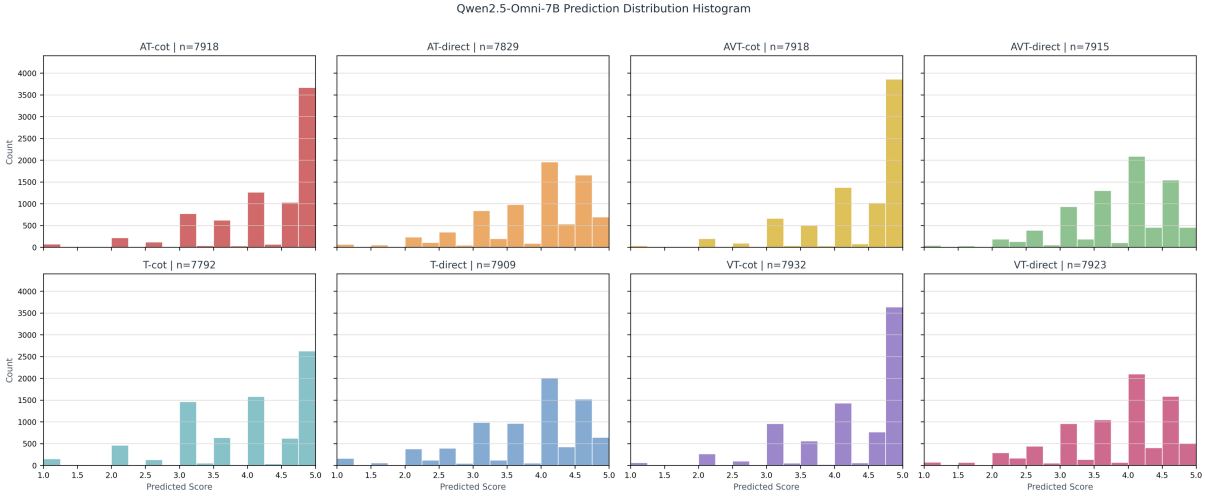
D.4 Prediction Histograms



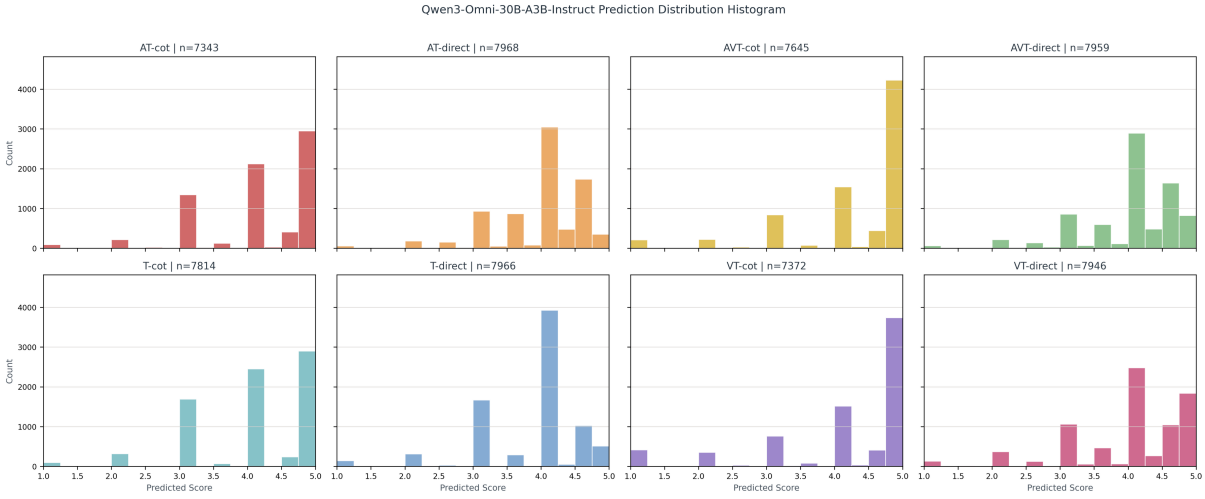
Qwen2-Audio-7B-Instruct



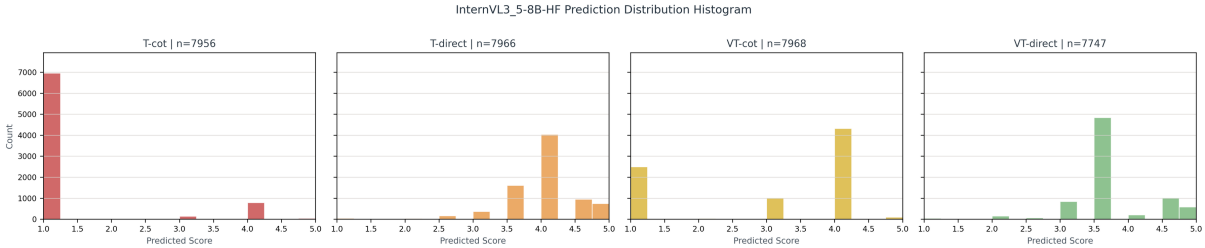
Qwen2.5-VL-7B-Instruct



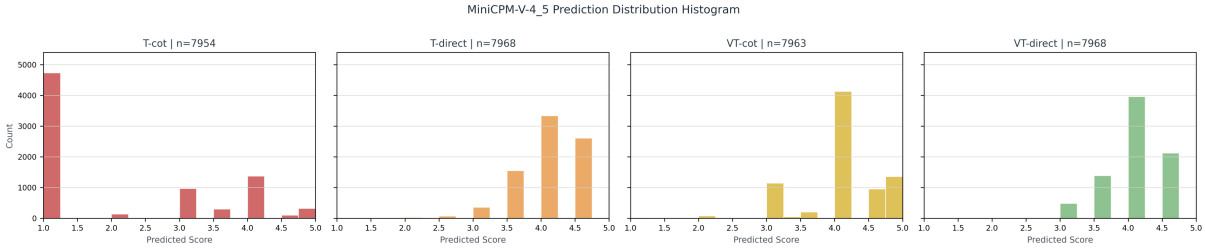
Qwen2.5-Omni-7B



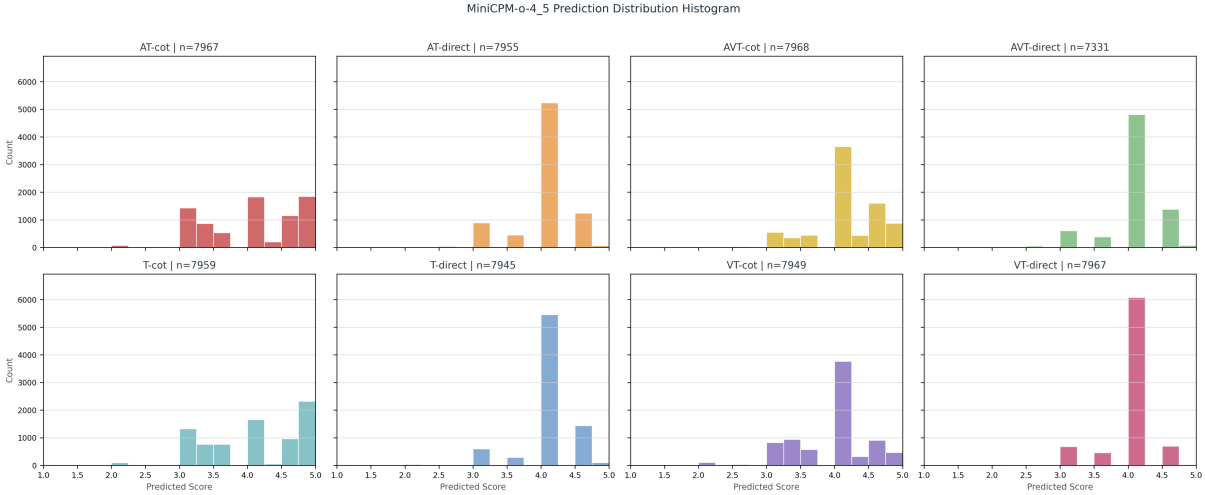
Qwen3-Omni-30B-A3B-Instruct



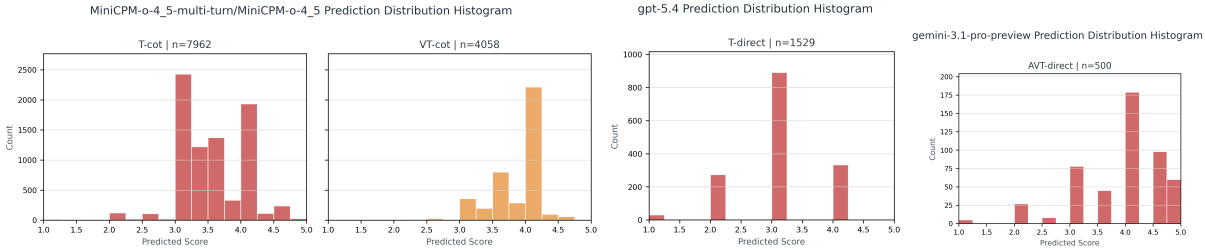
InternVL3.5-8B-HF



MiniCPM-V-4.5



MiniCPM-o-4.5



MiniCPM-o-4.5 multi-turn GPT-5.4 Gemini-3.1-Pro-Preview

Figure 15: **Prediction histograms.**