A Multi-To-One Interview Paradigm for Efficient MLLM Evaluation

Ye Shen[1,2], Junying Wang[2,3], Farong Wen[1,2], Yijin Guo[1,2],
Qi Jia[2], Zicheng Zhang[2,†], Guangtao Zhai[1,2,†]
[1]Shanghai Jiao Tong University, [2]Shanghai Artificial Intelligence Laboratory,
[3]Fudan University

***Abstract*—The rapid progress of Multi-Modal Large Language Models (MLLMs) has ushered in an explosion of benchmarks to evaluate and rank their capabilities. However, traditional full-coverage Question and Answering (Q&A) evaluation strategies suffer from high redundancy and low efficiency. Drawing inspiration from real-world human interview protocols, we propose the multi-to-one interview paradigm to break the limitations of conventional MLLM assessment. Our paradigm comprises three synergistic components, which are (i) a two-stage interview strategy with a pre-interview for initial proficiency calibration followed by a formal inquiry phase, (ii) dynamic adjustment of interviewer weights, which adaptively calibrates the influence of multiple evaluator models to mitigate individual biases and ensure a fair assessment, and (iii) an adaptive difficulty mechanism that iteratively selects questions based on real-time performance to probe the model's capability boundaries. Ablation experiments demonstrate that each component plays a role in the overall paradigm. Extensive experiments across diverse benchmarks demonstrate that the proposed paradigm achieves significantly higher correlation than random sampling, with improvements of up to 17.6% in PLCC and 16.7% in SRCC, while reducing the number of required questions. These findings demonstrate that our paradigm provides an adaptive, reliable and efficient alternative for large-scale MLLM benchmarking.**

***Index Terms*—MLLM Evaluation, Multi-To-One Interview, Efficiency**

I. Introduction

MLLMs have catalyzed significant breakthroughs across a spectrum of modalities, including images, videos, audio, and 3D content [1], [2]. However, this burgeoning capability brings an inherent paradox that as models advance rapidly, accurately and efficiently distilling and differentiating their core competencies becomes increasingly intricate [3]. While the proliferation of benchmarks has provided a vast landscape of evaluation tasks, ranging from foundational spatial perception to high-level cognitive reasoning [4], in essence, the sheer quantity of available test instances no longer serves as a reliable proxy for the quality of the assessment [5].

Conventional evaluation frameworks [6], [7] typically rely on a static approach, employing either exhaustive Q&A on entire datasets or random sampling of subsets. Under this regime, MLLMs labor through fixed, pre-determined test papers that fail to adapt to its specific strengths and weaknesses. Traditional paradigm is plagued by severe redundancy, that is to say, a significant portion of test instances are highly correlated in terms of the underlying knowledge required [8]. This inefficiency underscores a fundamental realization that fewer instances are sufficient for reliable ranking than for exhaustive evaluation. There is, therefore, a pressing need for a paradigm shift to a more agile, efficient, and strategic evaluation mechanism that can dynamically navigate the model's capability boundaries.

To transcend the rigid constraints of conventional testing, we propose a fundamental reconceptualization of the evaluation framework, shifting from passive examination to dynamic interaction. This transition is deeply inspired by the real-world mechanics of human recruitment, where rather than subjecting a candidate to a monolithic and exhaustive list of questions, a panel of evaluators utilizes a condensed set of strategically calibrated queries to probe the candidate's depth of knowledge and cognitive agility. Human **multi-to-one interview** achieves high efficiency while taking accuracy into account.

Consequently, we introduce a novel interview paradigm to address the inefficiency of conventional evaluation strategies. Specifically, our approach consists of three key components: (i) a **two-stage interview strategy**, including a lightweight pre-interview for initial difficulty calibration and a formal interview for comprehensive capability assessment; (ii) **dynamic adjustment of interviewer weights**, allowing diverse models to evaluate the interviewee more fairly and thoroughly and avoid individual bias; and (iii) an **adaptive difficulty mechanism** that updates subsequent questions based on both the current round's difficulty and the interviewee's performance, ensuring broad coverage of capability levels. Together, these synergistic components enable **comprehensive, accurate, fair, and efficient** evaluation of MLLMs.

Our main contributions are as follows:

- We propose a **multi-to-one interview paradigm** for MLLM evaluation, incorporating (i) a two-stage interview strategy, (ii) dynamic interviewer weight adjustment, and (iii) an adaptive difficulty mechanism.
- We demonstrate that this paradigm provides **reliable, fair, and efficient reflections** of MLLM capabilities, capturing both overall performance and difficulty-aware distributions.
- Extensive experiments on MMT-Bench, ScienceQA, and SEED-Bench show that our approach consistently outperforms random sampling and achieves up to **17.6% PLCC and 16.7% SRCC improvements** over full-coverage Q&A testing with fewer questions.

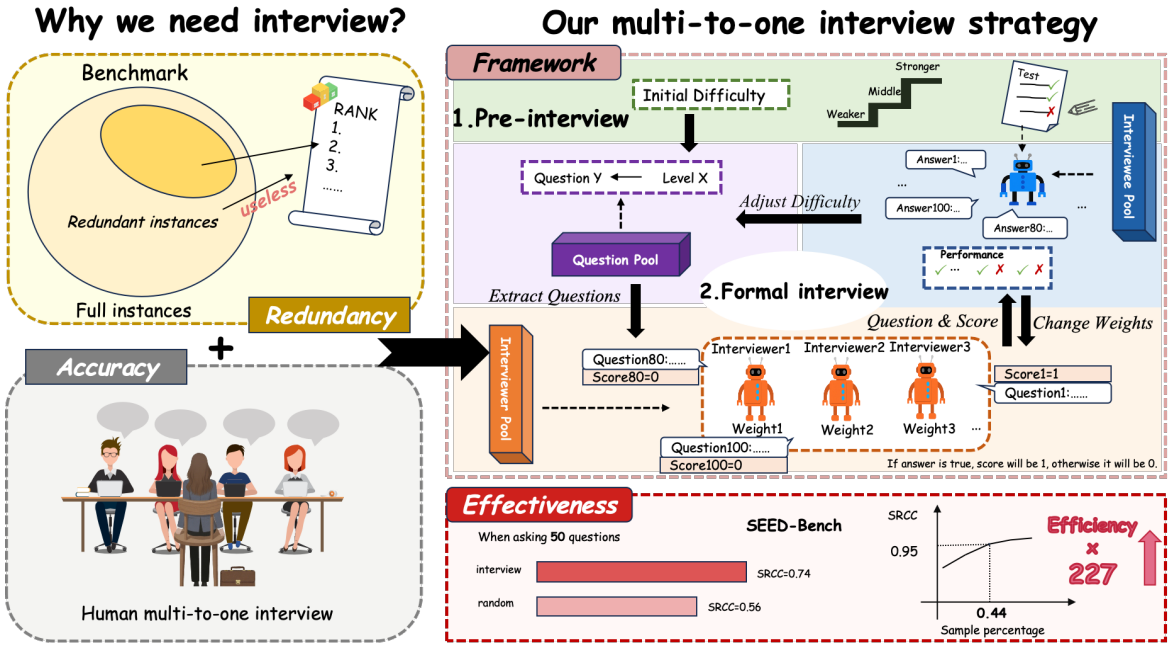


Fig. 1. Using full instances in the benchmark is less efficient for evaluation. Considering the efficiency of human multi-to-one interviews, we propose a multi-to-one interview paradigm in evaluating the capabilities of models and demonstrate its efficiency, accuracy, and reliability.

II. Related Work

A. Multi-Model Large Language Models

Large Language Models (LLMs) have exhibited exceptional proficiency in text processing and reasoning [9]. By integrating these strengths, Multi-Modal Large Language Models are further empowered to tackle intricate multi-modal challenges [9]. Seminal works like CLIP-ViT [10] established the baseline for visual-textual alignment. Subsequently, advanced models, including Claude-4-Sonnet [11], GPT-4o [12], Grok-4 [13], and Gemini-2.5-Pro [14], have continually elevated performance standards, effectively handling complex tasks in real-world environments. While these models exhibit excellent capabilities, the rapid evolution of their underlying architectures has outpaced the development of static benchmarks, creating a gap in comprehensively and quantitatively gauging their multifaceted boundaries [15]. Therefore, how to evaluate these models has emerged as a pivotal imperative.

B. MLLM Benchmarks

The meteoric rise of MLLMs has necessitated a paradigm shift in evaluation methodologies, driving the transition from single-task metrics to multi-dimensional, comprehensive assessments [9]. While foundational benchmarks like VQA [16] were pivotal in evaluating basic visual recognition, they fell short in capturing the advanced cognitive and complex reasoning capabilities inherent in modern foundation models. Consequently, researchers have pivoted towards designing holistic evaluation frameworks tailored to MLLMs' evolving characteristics [17]–[19]. For instance, SEED-Bench [20] significantly expands the dimensionality of evaluation, covering varied modalities and granularities, while MMT-Bench [7] further extends the scope to encompass intricate tasks embedded within realistic application scenarios. Despite this progress in benchmark breadth, the prevailing testing paradigm remains rigid. Most frameworks rely on a static, full-coverage Q&A strategy, where models are indiscriminately tested on the entire dataset [21]. This approach inevitably leads to computational redundancy and a critical lack of adaptability to the model's specific proficiency level. Therefore, a **more effective, dynamic, and efficient** evaluation strategy is urgently required to maximize assessment precision while minimizing overhead.

III. Interview Framework

In this section, we introduce the well-designed interview framework, combining (i) a two-stage interview strategy, (ii) dynamic adjustment of interviewer weights, and (iii) adaptive difficulty mechanism. Simultaneously, we propose computational methods to verify the effectiveness.

A. The two-stage interview strategy

The complete interview process includes a pre-interview for the coarse-grained assessment of the abilities of interviewee models from interviewee pool ($\mathcal{I}$), and a formal interview for a fine-grained evaluation to determine the models' comprehensive performance accurately.

Firstly, the pre-interview phase utilizes a written test of $middle$ difficulty to initially assess the model's capabilities. In specific, j $middle$-level questions are randomly selected from the corresponding question pool to test the model. Then, accuracy is calculated to determine the initial difficulty of the

formal interview, which can be expressed in formula 1. In this way, we can roughly determine the level of ability.

$$L_i = \begin{cases} middle + 1 & \text{if } acc > \beta, \\ middle & \text{if } acc = \beta, \\ middle - 1 & \text{otherwise,} \end{cases} \tag{1}$$

where L_i represents the initial difficulty level, acc represents interviewee's test results of the pre-interview, β represents an adjustable accuracy threshold.

Secondly, the formal interview is split into multiple rounds, each of which includes T questions. During each query, the selected interviewer model W is responsible for extracting multiple classical questions from categories arranged in stack matching the current difficulty, and judging whether the interviewee's responses are correct. The interviewee is required to answer the query. Then according to the interviewee's performance in the difficulty level of this round, the difficulty level of questions in the next round will be adjusted.

The whole interview will end once reaching the set number of questions N.

B. Dynamic adjustment of interviewer weights

To enhance the reliability and fairness of the proposed interview paradigm, we introduce a dynamic weighting mechanism for interviewer models. In brief, m interviewers are sampled from the interviewer pool $\mathcal{E}$ and assigned equal initial weights. During the iterative Q&A process, the weights of each interviewer, denoted as $w_i{}_{i=1}^{m}$, are updated after each query based on the historical responses generated by the interviewee model under their supervision. Weights determine the probability of selecting each interviewer to pose subsequent questions. The weight update rule is defined based on formula 2.

$$w_i^{t+1} = \begin{cases} (1-\alpha) \cdot w_i^t, & \text{if } acc_i = 0 \text{ or } acc_i = 1, \\ (1+\alpha) \cdot w_i^t, & \text{otherwise,} \end{cases} \tag{2}$$

where i represents the i-th interviewer model, t represents the t-th Q&A, acc_i represents performance of the interviewee under the inspection of the i-th interviewer. The hyperparameter α controls the magnitude of weight adjustment. To prevent extreme bias toward any interviewer, each weight w_i is constrained within the interval $[0.5, 2]$.

Intuitively, interviewers that consistently elicit either completely correct or completely incorrect responses are down-weighted, as such outcomes may indicate overly easy or overly difficult questioning. In contrast, interviewers that produce intermediate accuracy are up-weighted, since their questions tend to be more discriminative and informative for capability estimation. This adaptive mechanism encourages a balanced and difficulty-based evaluation process, thereby improving both ranking stability and robustness.

C. Adaptive difficulty mechanism

We design an adaptive difficulty adjustment strategy to iteratively steer the evaluation process toward the capability limit of the models, thereby improving discriminative power while avoiding unnecessary sampling of overly easy or overly difficult questions.

Algorithm 1 Formal Interview Procedure

Input: L_i, L^r, $\{w_i\}_{i=1}^m$, T, N, P, $\mathcal{I}$, $\mathcal{E}$, $F(\cdot)$, $G(\cdot)$, $H(\cdot)$
Output: Final Score S

```
Target at each interviewee in I
Randomly select m interviewers from E
for t = 1 to N do                                  ▷ Select interviewer model
    W ← {w_i^t}_{i=1}^m
    Selected interviewer extract questions of L_i
    for r = 1 to ⌊N/T⌋ do                          ▷ Adjust round difficulty
        L^{r+1} ← F(L^r, acc^r)
        if L^{r-k} = L^r and L^{r-k+1} = L^{r+1} for k = 1, 2, ..., P then
            L^{r+1} ← G(L^{r-1}, n)
        end if
        if no samples exist in L^r then
            L^{r+1} ← H(L^r, acc^r)
        end if
    end for
end for
Calculate accuracy of all {L_r}_{r=1}^{10}
return S
```

Specifically, after r-th round, accuracy of the interviewee model at difficulty level L^r is computed. The difficulty of the next round L_{r+1} is then updated according to formula 3:

$$L^{r+1} = F(L^r, acc^r) = \begin{cases} L^r + 1, & \text{if } acc^r > \beta, \\ L^r, & \text{if } acc^r = \beta, \\ L^r - 1, & \text{if } acc^r < \beta, \end{cases} \tag{3}$$

where acc^r represents the model's performance at r-th level, and β is a predefined accuracy threshold. This mechanism enables the evaluation process to gradually converge toward the model's effective capability boundary.

To prevent oscillatory behavior caused by repeated difficulty switching between adjacent levels, we introduce an additional adjustment rule. When the alternating pattern between L^{r-1} and L^r repeats for P consecutive rounds, the difficulty is updated using formula 4:

$$L^{r+1} = G(L^{r-1}, n) = \begin{cases} L^{r-1} + k, & \text{if } L^{r-1} > n, \\ L^{r-1} - k, & \text{otherwise,} \end{cases} \tag{4}$$

where n is an adjustable difficulty threshold and k is a predefined step size. This rule enables a controlled jump across difficulty levels, ensuring broader coverage of the difficulty spectrum and preventing the interview process from being trapped in a narrow local region.

Furthermore, when no questions are available at difficulty level L^r, we adjust the difficulty according to:

$$L^{r+1} = H(L^r, acc^r) = \begin{cases} L^r + 1, & \text{if } acc^r > \beta, \\ L^r - 1, & \text{otherwise,} \end{cases} \tag{5}$$

TABLE I
A COMPARISON BETWEEN THE MULTI-TO-ONE INTERVIEW PARADIGM AND THE RANDOM STRATEGY, WHERE AVG.IMPROVEMENT MEANS AVERAGE ABSOLUTE INCREASE (PERCENTAGE POINTS) OF OUR PARADIGM OVER RANDOM SAMPLING IN EACH METRIC.

Benchmark	**MMT-Bench**			**ScienceQA**			**SEED-Bench**		
Number	**SRCC**	**PLCC**	**KRCC**	**SRCC**	**PLCC**	**KRCC**	**SRCC**	**PLCC**	**KRCC**
Random Strategy: Questions picked randomly									
20	0.5573	0.6654	0.4279	0.3515	0.4205	0.2284	0.4277	0.4391	0.3090
30	0.5687	0.6705	0.4377	0.3668	0.4316	0.2433	0.4590	0.4876	0.3283
50	0.6397	0.7021	0.4955	0.4688	0.6183	0.3512	0.5577	0.5894	0.4085
80	0.6496	0.6985	0.5373	0.5592	0.6376	0.4345	0.6129	0.6925	0.5091
100	0.6932	0.6470	0.5426	0.5984	0.6402	0.4556	0.6554	0.6907	0.4820
Multi-to-One Interview Paradigm (proposed): Questions picked during interview									
20	0.6202	0.7275	0.4273	0.5911	0.7255	0.4366	0.5958	0.6780	0.4372
30	0.6424	0.7608	0.4557	0.6289	0.7225	0.4956	0.7122	0.6579	0.5494
50	0.6749	0.7957	0.5148	0.6348	0.7262	0.5118	0.7377	0.7110	0.5733
80	0.7091	0.8303	0.5385	0.6435	0.7213	0.4882	0.7491	0.7239	0.5852
100	0.7109	0.8308	0.5503	0.6547	0.7328	0.5013	0.7532	0.7251	0.5962
Avg. improvement	**4.98%**	**11.23%**	**0.91%**	**16.17%**	**17.60%**	**14.41%**	**16.71%**	**11.93%**	**14.09%**

where β is the same adjustable threshold in formula 3. This fallback mechanism ensures continuity of the interview process and facilitates finer-grained capability estimation around the current difficulty boundary.

D. Evaluation

We define models' capability level $L*$ as the max level where the model achieves an accuracy not less than 50% within a small tolerance θ.

$$L* = max\{L|acc_L \geq 0.5 - \theta\}, \tag{6}$$

where acc_L represents the accuracy of model's responses in the level L. Based on $L*$, we can calculate the final score S corresponding to each interviewee model.

$$S = 0.1L* + 0.1\frac{L* \cdot acc_{L*} + (L*+1) \cdot acc_{L*+1}}{2L*+1}, \tag{7}$$

However, if there is no $acc_L \geq 0.5 - \theta$, we consider $L* = 1$, then $S = 0.1 + 0.1\frac{acc_1 + 2 \cdot acc_2}{3}$, and if $L* = 10$, we control the score S to be 1.

According to the formula 7 and other situations, we can estimate the final capability level of models for reliable rankings. The performance of MLLMs in full-coverage Q&A testing serves as the ground truth. 3 widely adopted metrics is utilized for correlation analysis, which are Spearman's Rank Correlation Coefficient (SRCC) [22], Pearson's Linear Correlation Coefficient (PLCC) [23], and Kendall's Rank Correlation Coefficient (KRCC) [24].

IV. EXPERIMENT

This section specifies the relevant prepare work and the settings of the experiments.

A. Prepare work

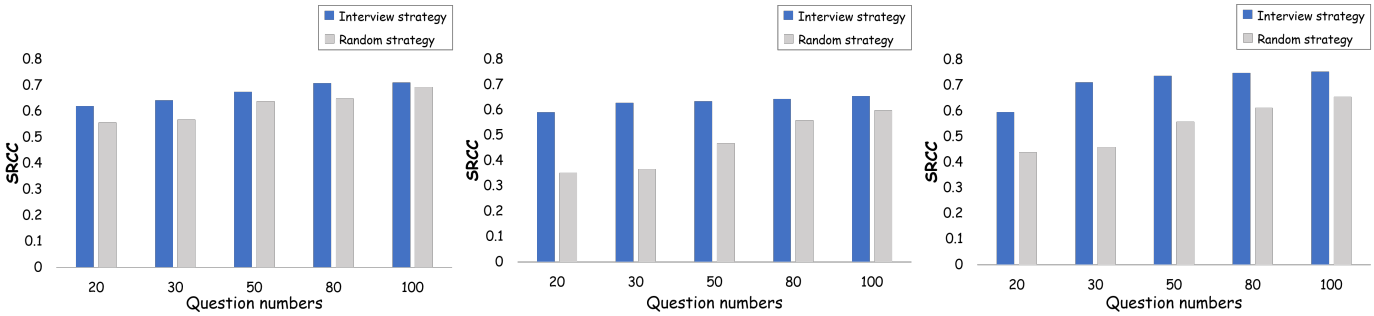
We construct a dataset with clear difficulty levels and introduce the models involved in the experiments.

TABLE II
MLLM INTERVIEWEES AND MLLM INTERVIEWERS ILLUSTRATION.

Calling Method	**Model**
API Call	GPT-4.1-Nano [25], GPT-4o-Mini [12], GPT-4o [12], Grok-2 [26], Grok-4 [13], Claude-3.5-Sonnet [27], Claude-4-Sonnet [11], Qwen-VL-Plus [28], Gemini-2.0-Flash [14], Gemini-2.5-Pro [14], Gemini-2.5-Flash [14]
Local Call	Phi-4-mini-Instruct [29], Qwen2.5-VL-7b-Instruct [30], Qwen2.5-VL-72B-Instruct [30], Qwen2.5-VL-32B-Instruct [30], Internlm3-8B-Instruct [31], Internlm2.5-20B-Chat [31], Gemma-3-27B-It [32], Llama-3.2-11B-Vision-Instruct [33]

1) Related benchmarks: Traditional evaluation questions often lack a systematic difficulty division, making it difficult to efficiently assess interviewees' abilities across different fields. To construct specific benchmarks with difficulty attributes, 10 typical models are selected to examine the difficulty of questions, which include GPT-4o [12], Deepseek-VL [34], Qwen-2.5-VL [30], Gemini-1.5-pro [35], Grok-2 [26], Kimi-VL [36], Phi-3 [29], Claude3-7 [27], HunYuan [37] and InternVL3 [38]. Specifically, the difficulty level of each question is obtained through subtracting the number of models that answered correctly from 11. If level is calculated as 11, it is treated as 10. Ultimately, questions in MMT-Bench [7], ScienceQA [6] and SEED-Bench [20] are successfully divided into 10 difficulty levels.

2) Models: The interviewer and interviewee models used in the experiments are shown in Tab. II, of which eleven are closed-source and eight are open-source. The interviewer pool contains the eleven closed-source models, while the interviewee pool contains all the models.



(a) MMT-Bench (b) ScienceQA (c) SEED-Bench

Fig. 2. Overview of SRCC performance corresponding the question numbers on 3 Benchmarks clearly distinguishes the gap between the two methods.

B. Settings

Each model in Tab. II is designated as an interviewee, with $m = 3$ randomly selected models via API call as interviewers. The pre-interview administers $j = 3$ random questions of $middle = 5$. Formal interview rounds each include $T = 3$ questions. Interviewers start with weight of 1.0. Parameters are set as follows, including $\alpha = 0.2$ to ensure the rationality of weight adjustment, $\beta = 0.5$ to accurately measure model level, and $n = 7$ to cover full difficulties, $k = 3$ to cover as more levels as possible, and $P = 3$, $N = 200$.

Accuracy across different difficulties are calculated to verify interviewees' capability distribution and set up a control group where we randomly select questions from benchmarks.

V. Analysis

The effectiveness of our paradigm is demonstrated across three benchmarks: MMT-Bench, ScienceQA, and SEED-Bench, compared with the random strategy.

A. Accuracy of our paradigm

According to the results reported in Tab. I, the proposed multi-to-one interview paradigm consistently achieves higher SRCC, PLCC, and KRCC scores than random sampling across most question settings. This indicates that our paradigm produces evaluation results that are more strongly correlated with the reference rankings. The performance gap is particularly noticeable on SEED-Bench, which contains a large number of questions with diverse difficulty levels. Such results suggest that random sampling may fail to capture representative difficulty distributions under limited questions, while our interview selection strategy can more effectively identify informative questions. Consequently, the multi-to-one interview paradigm enables more accurate capability assessment, even when the number of available evaluation questions is constrained.

B. Efficiency of our paradigm

The SRCC trends illustrated in Fig. 2 further demonstrate the superior sample efficiency of our proposed paradigm across all three benchmarks. Compared with random sampling, our strategy consistently maintains higher rank correlation as the number of selected questions increases. The advantage is particularly pronounced in the low-budget regime. For example, when only 30 questions are selected on ScienceQA, the proposed paradigm achieves an SRCC of 0.6289, whereas random sampling attains merely 0.3668, corresponding to a 71.46% relative improvement.

This substantial gap suggests that random sampling struggles to preserve reliable and accurate model rankings under limited evaluation budgets. In contrast, the interview-based selection mechanism prioritizes more informative and discriminative questions at early stages, leading to faster convergence towards stable ranking performance. As the number of questions increases, the gap gradually narrows, indicating that random sampling requires substantially more samples to achieve comparable reliability.

TABLE III
Ablation study on different strategies.

Pre-interview	Interviewer Weight	Adaptive Difficulty	SRCC
×	✓	✓	0.7029
✓	×	✓	0.6834
✓	✓	×	0.4357
✓	✓	✓	**0.7532**

C. Ablation Experiment

We conduct component-level ablation studies on SEED-Bench and report the maximum SRCC achieved under each setting to evaluate the contribution of individual modules.

As shown in Tab. III, removing any component consistently results in a noticeable decline in SRCC, demonstrating that each module contributes positively to the overall effectiveness of the proposed framework. This consistent degradation also suggests that the improvements do not stem from a single dominant factor, but rather from the coordinated interaction of multiple components.

Notably, the removal of the adaptive difficulty mechanism leads to the most significant performance drop. This observation highlights the critical role of adaptive difficulty mechanism in reliable capability estimation. By adaptively steering the interview process toward appropriately challenging questions, the mechanism helps maintain discriminative power

across models of varying strengths. Without it, the evaluation may suffer from either overly easy or overly difficult question distributions, resulting in less stable, less accurate, and less efficient ranking correlations.

VI. CONCLUSION

In summary, we propose a multi-to-one interview paradigm that substantially improves the efficiency of MLLM evaluation while offering a new perspective on conventional benchmarking methodologies. By reformulating evaluation as an adaptive interview process rather than static question sampling, our framework enables more accurate capability estimation under limited questions and enhances ranking consistency across models. The empirical results across multiple benchmarks further validate its effectiveness and robustness.

However, several limitations remain. Firstly, the paradigm is still built upon Q&A evaluation, which may not fully capture complex reasoning processes and open-ended generation capabilities. Secondly, the current interviewer selection strategy is relatively simple and may not fully exploit the diversity of potential evaluators. In the future, this framework could be extended to support automated benchmark construction and further explored for its potential in cross-lingual evaluation.

VII. ACKNOWLEDGMENTS

Large language models, including GPT-5.1 and Gemini-3-Pro, were used solely for text refinement and drafting code comments, and all outputs were carefully reviewed, edited, and verified by the authors. This work was supported in part by the National Natural Science Foundation of China under Grants 62225112, 62132006, U24A20220 and was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0124104) in collaboration with Shanghai Artificial Intelligence Laboratory.

REFERENCES

[1] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, Nov. 2024. [Online]. Available: http://dx.doi.org/10.1093/nsr/nwae403

[2] Z. Zhang, J. Wang, F. Wen, Y. Guo *et al.*, "Large multimodal models evaluation: A survey," *SCIENCE CHINA Information Sciences*, vol. 68, no. 12, pp. 221 301–221 369, 2025.

[3] J. Huang and J. Zhang, "A survey on evaluation of multimodal large language models," 2024. [Online]. Available: https://arxiv.org/abs/2408.15769

[4] Z. Huang, Z. Wang, S. Xia, X. Li, H. Zou, R. Xu *et al.*, "Olympicarena: Benchmarking multi-discipline cognitive reasoning for superintelligent ai," *Advances in Neural Information Processing Systems*, vol. 37, pp. 19 209–19 253, 2024.

[5] P. Liang *et al.*, "Holistic evaluation of language models," 2023. [Online]. Available: https://arxiv.org/abs/2211.09110

[6] T. Saikh, T. Ghosal, A. Mittal, A. Ekbal, and P. Bhattacharyya, "Scienceqa: A novel resource for question answering on scholarly articles," *International Journal on Digital Libraries*, vol. 23, no. 3, pp. 289–301, 2022.

[7] K. Ying, F. Meng, J. Wang, Z. Li, H. Lin, Y. Yang *et al.*, "Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi," 2024. [Online]. Available: https://arxiv.org/abs/2404.16006

[8] Z. Zhang, X. Zhao, X. Fang, C. Li, X. Liu, X. Min *et al.*, "Redundancy principles for mllms benchmarks," 2025. [Online]. Available: https://arxiv.org/abs/2501.13953

[9] C. Fu, Y.-F. Zhang, S. Yin, B. Li, X. Fang, S. Zhao *et al.*, "Mme-survey: A comprehensive survey on evaluation of multimodal llms," 2024. [Online]. Available: https://arxiv.org/abs/2411.15296

[10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[11] Anthropic, "Introducing claude 4," https://www.anthropic.com/news/claude-4, May 2025.

[12] OpenAI, "Hello GPT-4o," https://openai.com/index/hello-gpt-4o/, 2024.

[13] xAI, "Grok 4," https://x.ai/news/grok-4, July 2025.

[14] G. Team *et al.*, "Gemini: a family of highly capable multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2312.11805

[15] R. Burnell *et al.*, "Rethink reporting of evaluation results in ai," *Science (New York, N.Y.)*, vol. 380, pp. 136–138, 04 2023.

[16] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2425–2433.

[17] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin *et al.*, "Mme: A comprehensive evaluation benchmark for multimodal large language models," 2025.

[18] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao *et al.*, "Mmbench: Is your multi-modal model an all-around player?" 2024. [Online]. Available: https://arxiv.org/abs/2307.06281

[19] X. Yue *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 9556–9567.

[20] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, "Seed-bench: Benchmarking multimodal llms with generative comprehension," 2023. [Online]. Available: https://arxiv.org/abs/2307.16125

[21] F. M. Polo, L. Weber, L. Choshen, Y. Sun, G. Xu, and M. Yurochkin, "tinybenchmarks: evaluating llms with fewer examples," 2024. [Online]. Available: https://arxiv.org/abs/2402.14992

[22] C. Spearman, "The proof and measurement of association between two things." 1961.

[23] K. Pearson, "Vii. note on regression and inheritance in the case of two parents," *proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.

[24] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1-2, pp. 81–93, 1938.

[25] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman *et al.*, "Gpt-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2303.08774

[26] xAI, "Grok-2 beta release," https://x.ai/news/grok-2, 2024.

[27] A. Anthropic, "The claude 3 model family: Opus, sonnet, haiku," *Claude-3 Model Card*, vol. 1, no. 1, p. 4, 2024.

[28] J. Bai *et al.*, "Qwen technical report," 2023. [Online]. Available: https://arxiv.org/abs/2309.16609

[29] M. Abdin *et al.*, "Phi-4 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2412.08905

[30] S. Bai *et al.*, "Qwen2. 5-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923

[31] Z. Cai *et al.*, "Internlm2 technical report," 2024. [Online]. Available: https://arxiv.org/abs/2403.17297

[32] G. Team *et al.*, "Gemma 3 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2503.19786

[33] A. Grattafiori *et al.*, "The llama 3 herd of models," 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[34] Z. Wu *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," 2024. [Online]. Available: https://arxiv.org/abs/2412.10302

[35] G. Team *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 2024. [Online]. Available: https://arxiv.org/abs/2403.05530

[36] K. Team *et al.*, "Kimi-vl technical report," 2025. [Online]. Available: https://arxiv.org/abs/2504.07491

[37] X. Sun *et al.*, "Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent," 2024. [Online]. Available: https://arxiv.org/abs/2411.02265

[38] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 24 185–24 198.