INTERMT: Multi-Turn Interleaved Preference Alignment with Human Feedback

Boyuan Chen[1*] **Donghai Hong**[1*] **Jiaming Ji**[1*] **Jiacheng Zheng**[2]
Bowen Dong[1] **Jiayi Zhou**[1] **Kaile Wang**[1] **Josef Dai**[1] **Xuyao Wang**[1]
Wenqi Chen[1] **Qirui Zheng**[1] **Wenxin Li**[1] **Sirui Han**[2] **Yike Guo**[2] **Yaodong Yang**[1†]

[1] Peking University [2] Hong Kong University of Science and Technology

Abstract

As multimodal large models (MLLMs) continue to advance across challenging tasks, a key question emerges: ***What essential capabilities are still missing?*** A critical aspect of human learning is continuous interaction with the environment – not limited to language, but also involving multimodal understanding and generation. To move closer to human-level intelligence, models must similarly support **multi-turn**, **multimodal interaction**. In particular, they should comprehend interleaved multimodal contexts and respond coherently in ongoing exchanges. In this work, we present **an initial exploration** through the INTERMT – **the first preference dataset for *multi-turn* multimodal interaction**, grounded in real human feedback. In this exploration, we particularly emphasize the importance of human oversight, introducing expert annotations to guide the process, motivated by the fact that current MLLMs lack such complex interactive capabilities. INTERMT captures human preferences at both global and local levels into nine sub-dimensions, consists of 15.6k prompts, 52.6k multi-turn dialogue instances, and 32.4k human-labeled preference pairs. To compensate for the lack of capability for multi-modal understanding and generation, we introduce an agentic workflow that leverages tool-augmented MLLMs to construct multi-turn QA instances. To further this goal, we introduce INTERMT-BENCH to assess the ability of MLLMs in assisting judges with multi-turn, multimodal tasks. We demonstrate the utility of INTERMT through applications such as judge moderation and further reveal the *multi-turn scaling law* of judge model. We hope the open-source of our data can help facilitate further research on aligning current MLLMs to the next step.

1 Introduction

Humans perceive the world through dynamic, multimodal interactions involving text, images, audio, video, and more [1, 2, 3]. Building on the success multimodal large language models (MLLMs) [4, 5, 6, 7, 8], recent efforts aim to develop general-purpose AI assistants that handle multiple mixed modalities [9, 10, 11]. A key feature of such general-purpose assistants is to engage in natural *multi-turn* conversations, perceive and generate any modality (*i.e.*, *interleaved multimodal understanding and generation*), to enable more smooth interaction and grounded understanding [9, 12, 11, 13, 14].

Recent years have seen community efforts in transplanting alignment techniques (*e.g.*, Reinforcement Learning from Human Feedback (RLHF)) from the text modality [15, 8, 16, 17] to multiple modalities settings [13, 18, 19, 20, 21, 22, 14]. Within this line of research, most studies focus exclusively on either understanding [23, 18] or generation [22, 21]. The lack of alignment considerations for multimodal mixed input-output settings exacerbates the imbalance across modalities, *i.e.*, *modality*

[*]Equal contribution, [†]corresponding author. Project website: https://pku-intermt.github.io.

Preprint. Under review.

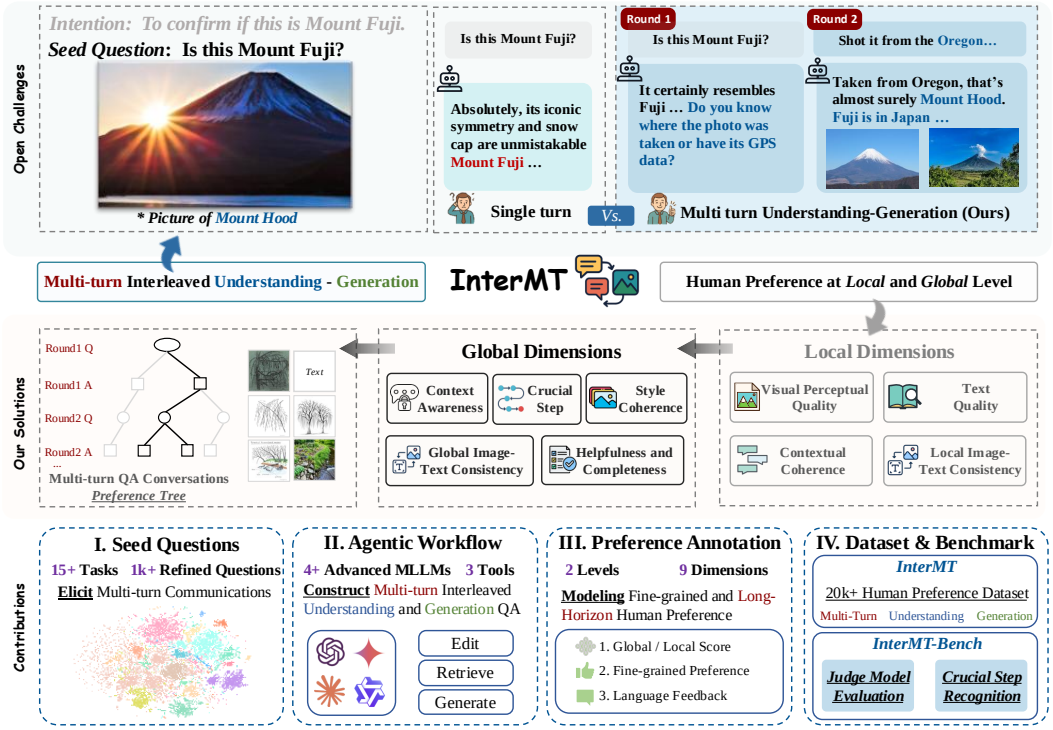

Figure 1: Motivated by the challenges of single-turn interactions in aligning with human intent, and the goal of constructing general-purpose AI assistants, we introduce **INTERMT**: **Data:** INTERMT, the first human preference dataset focused on **multi-turn**, multimodal **understanding** and **generation**; **Decoupled Helpfulness:** capturing human feedback at both the *local* (turn-level) and *global* (conversation-level); **Evaluation:** evaluating the capabilities of MLLMs as judge models.

disequilibrium.[14]. Furthermore, existing methods primarily focus on single-turn interactions, where an LLM generates a response from a prompt and receives immediate alignment feedback. However, real-world interactions typically occur in long-horizon conversations (*e.g.*, over 5 turns) and often feature interleaved multimodal inputs and outputs [24, 25, 26].

*How to improve **multi-turn** interleaved **understanding**- **generation** alignment via human feedback?*

Our reflections highlight several key issues in the alignment of MLLMs:

- **Modality Fusion via Harmonizing Understanding and Generation.** To build general-purpose AI assistants, high-fidelity perception and understanding alone are not sufficient. The system should also support the selective generation of multimodal outputs (*e.g.*, images) to effectively communicate, instruct, or interact with users in a natural and contextually appropriate manner.
- **Modeling Long-Horizon, Interleaved Multimodal Interactions.** Real-world user–AI exchanges typically span many turns and interleave text, vision, and other modalities. Such interactions demand not only precise instruction following but also sustained attention and reasoning over an evolving context, approaching near-human in-context reasoning capabilities.
- **Dynamic Human-in-the-Loop Alignment.** In extended, multimodal interactions, user preferences continually evolve. For example, a user may first ask the assistant to draw a vase, then—after inspecting the rendered image—request that the vase be repositioned or restyled for greater emphasis. Capturing and aligning with these emergent, dynamic preferences calls for genuine, iterative human feedback throughout the interaction.

In response, we introduce INTERMT, a human preference dataset designed to capture the complexity and diversity of human intent in **multi-turn** settings. Specifically, INTERMT targets vision-language interaction scenarios involving interleaved **understanding** and **generation**. To model dynamic human preferences, INTERMT comprises 15604 seed questions that elicit multi-turn, multimodal

conversations spanning 15+ domains. Helpfulness is then decomposed into 9 sub-dimensions, capturing both global (conversation-level) and local (turn-level) aspects of human feedback.

Our key contributions are summarized as follows:

- **The First Multi-turn Interleaved Preference Dataset:** To the best of our knowledge, INTERMT is the first dataset that captures real human preferences for tasks involving **multi-turn** and *interleaved multimodal **understanding** and **generation***. It contains 15604 unique seed questions across diverse categories, 52.6k multi-turn interleaved vision-language QA instances, and 32,459 sets of multi-dimensional human preference annotations.
- **Agent-based Construction Workflow:** INTERMT employs a carefully designed agent-based multi-turn QA construction workflow that leverages strong MLLMs augmented with external tools (*e.g.*, image editing, generation and retrieval) to simulate high-quality real multi-turn interactions.
- **Decoupled Helpfulness in Multi-turn Multimodal Scenarios:** INTERMT decomposes the concept of helpfulness for *multi-turn interleaved multimodal understanding and generation* into two distinct levels: *local* (turn-level) and *global* (conversation-level). At the local level, helpfulness is assessed for each individual turn, while at the global level, helpfulness is evaluated across the entire conversation. Furthermore, INTERMT breaks down helpfulness into 9 specific dimensions (*e.g.*, *contextual coherence*, *image-text consistency*, etc.), allowing for a detailed and nuanced evaluation of multi-turn, multi-modal interactions.
- **Effective for Multi-turn Alignment:** Building on INTERMT, we investigate methods to *model long-horizon values* and *align dynamic human values*. Our findings reveal the phenomenon of preference transfer in multi-turn multimodal interactions, which facilitates preference modeling for predicting human judgments. Additionally, we identify a *scaling phenomenon* in multi-turn multimodal judge moderation (Section 4.1).
- **One More Thing** We introduce INTERMT-BENCH to evaluate the ability of MLLMs in assisting judges across multi-turn, multimodal tasks, encompassing three parts: *Scoring Evaluation*, *Pair Comparison*, and *Crucial Step Recognition* (Section 4.2). Despite strong reasoning capabilities, advanced MLLMs (*e.g.*, o4-mini [27]) fail to align with human values in judgment tasks. However, they show potential in identifying crucial steps in long-context scenarios.

For more details about the motivation of our work, please refer to Appendix A.

2 Dataset

Our core contribution is the introduction of a human preference dataset designed for **multi-turn**, multimodal **understanding** and **generation** tasks. This section outlines the dataset's composition, the collection of prompts and multi-turn QA instances, and human annotation process.

2.1 Dataset Composition

The INTERMT dataset includes: (1) carefully crafted *seed questions* for multi-turn, multimodal conversations, and (2) fine-grained human preference annotations at both local and global conversation levels. Inspired by theories from linguistics, human-computer interaction, and cognitive psychology [28, 29, 30, 31, 32], the seed questions are rigorously selected and refined to enable more faithful simulation of real-world *interleaved multimodal understanding and generation* and *multi-turn* tasks. We collect preference data through score evaluations and pairwise comparisons of multi-modal responses at each conversation turn, based on four sub-dimensions. Global conversation helpfulness is then evaluated via five sub-dimensions. Incorporating natural language feedback further improves annotation quality and alignment with human intent. The **Data Card** for INTERMT is as follow:

- INTERMT is built from a corpus of 100k image-text examples, comprising 72.1% from open-source vision-language datasets, 22.8% from web data, and 5.1% from human-written content. All prompts are refined following constitutional guidelines to improve multi-turn compatibility, resulting in 15604 unique seed questions, as shown in Figure 2.
- Each seed question is expanded via an agent-based multi-turn QA construction workflow, producing at least 8 multi-turn QA instances per prompt. After pruning and filtering, we obtain 52.6k high-quality multi-turn QA instances, with 41.92% containing five or more turns.

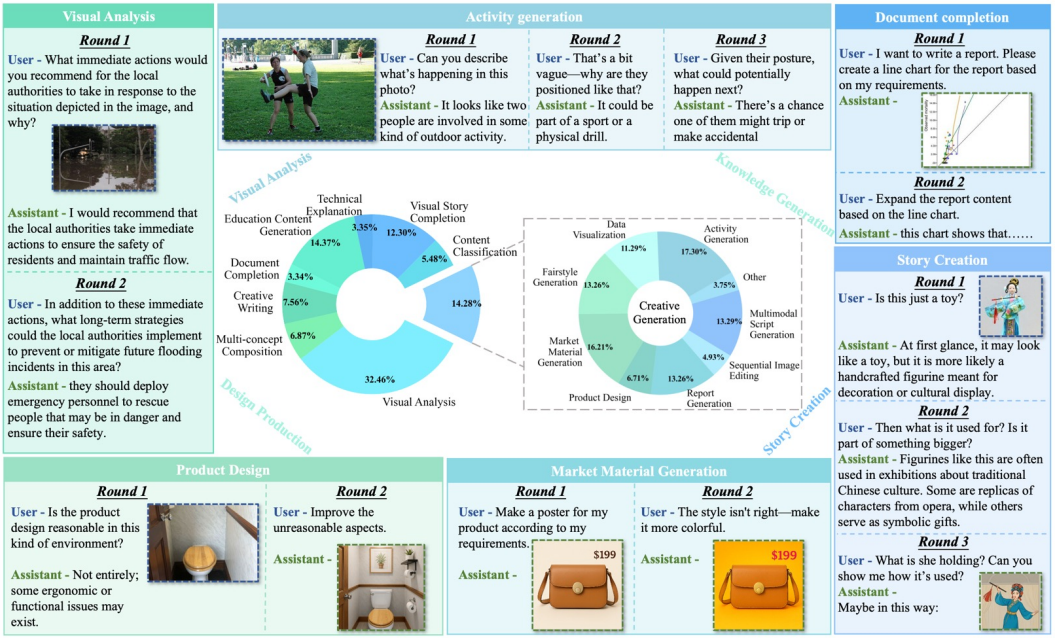

Figure 2: INTERMT includes over 15 tasks in vision-language scenarios, capturing communication examples across diverse **multi-turn** settings. These examples demonstrate **multi-turn**, interleaved **understanding** and **generation** in six representative domains.

- The resulting 52.6k QA instances cover 15+ vision-language **understanding** and **generation** tasks, such as image editing and visual tutorials. Each instance features interleaved textual and visual content in both inputs and outputs, with an average of 5.33 images per conversation.
- INTERMT features 32,459 human preference annotations, organized as score evaluation pairwise comparisons at both the local and global levels. Preferences are decomposed into 9 dimensions of helpfulness, accompanied by human-written critiques, refinement suggestions, and rationales.

2.2 Multi-turn QA Construction

Prompt Collection. INTERMT is constructed from 100k image-text QA instances collected from three primary sources: 72.1% from public datasets [14, 23, 33]; 22.8% from legally scraped web content; and the remaining 5.1% from researcher-curated, human-written prompts. These instances span diverse vision-language tasks, *e.g.*, activity generation, data visualization, and table analysis.

Drawing upon cognitive psychology theories [28, 29, 30, 31, 32], we identify five common scenarios that give rise to multi-turn conversations in real-world multimodal settings. Based on these scenarios, we filter, diversify, and rewrite the original image-text QA instances, resulting in 15604 unique *seed question*s. These questions serve as the initial round for generating multi-turn conversation data. Additional details can be found in Appendix D.

Tool-Augmented Agent Workflow for QA Construction. We identify two core challenges in constructing multi-turn QA instances that capture realistic scenarios of multimodal understanding and generation: (1) How to effectively simulate realistic human multi-turn conversations in multimodal contexts? (2) Given that current MLLMs lack interleaved understanding and generation capabilities [34, 24], how to construct interleaved QA instances that generalize across diverse real-world tasks?

To address these challenges, we propose a tool-augmented agent workflow that integrates powerful open-source and API-based models with image-centric tools. Within this framework, each agent simulates human-like conversations by either responding to the current query or generating follow-up questions based on the previous answer. Agents can invoke tools to generate, edit, or retrieve images, enabling the recursive construction of tree-structured, multi-turn interleaved image–text QA instances.

Agent Construction. The agent workflow is built upon a combination of strong open-source models [35, 36, 37, 4] alongside leading API-based models [38, 39, 40]. To support diverse mul-

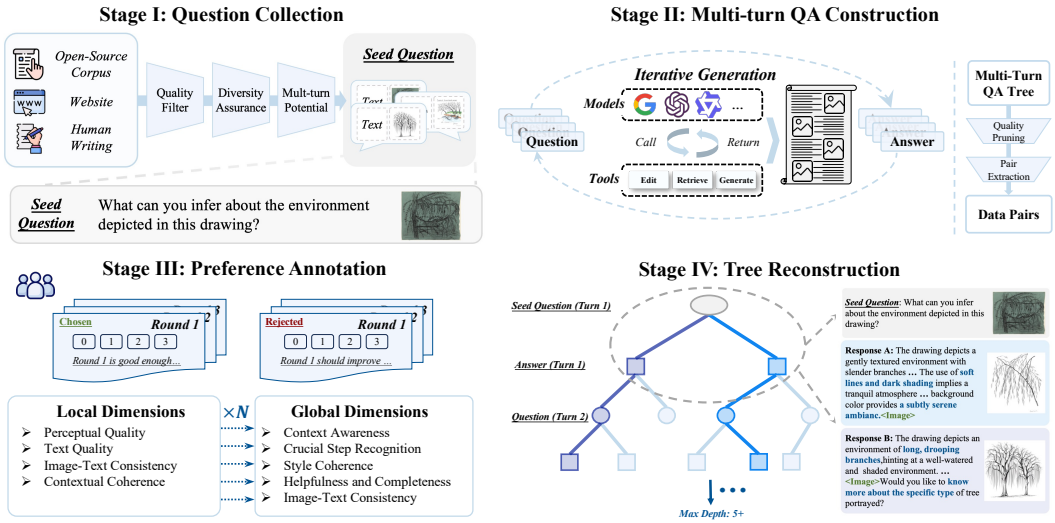

Figure 3: Overview of the four-stage pipeline for constructing INTERMT. **Stage I:** seed questions are harvested from open-source corpora, websites, and human writing, then filtered for perceptual quality, diversity, and multi-turn potential. **Stage II:** iterative calls to large models and external tools (*e.g.* edit, retrieve, generate) produce answer expansions and follow-up questions, forming a candidate QA tree. **Stage III:** human annotators perform per-turn (local) and conversation-level (global) evaluations—covering quality, coherence, context awareness, and completeness—to prune and select preferred branches. **Stage IV:** the retained branches are reassembled into deep, coherent QA trees (depth ≥ 5) yielding the final multi-turn QA pairs for model training.

timodal operations, three types of image-centric tools are integrated: (1) text-to-image generators (*e.g.*, FLUX.1-Schnell [41] and Stable-Diffusion [42]) for producing high-quality images based on prompts; (2) an image editing API (*e.g.*, Gemini-2.0-flash [43]) capable of cropping, highlighting, and modifying images; and (3) web-based retrieval interfaces for sourcing real-world visuals. During multi-turn QA generation, agents embed structured tokens such as `<Image,caption>` within the text to denote visual references after which GPT-4o [38] serves as a double classifier and verifier, automatically determining the appropriate tool call based on the image intent and context.

Iteratively Question and Response Generation. We begin with carefully crafted *seed questions* to initiate extended multimodal dialogues; at each turn, diverse agents generate a pool of 10 candidate follow-ups via a Socratic strategy, from which $\mathcal{M}$ (typically 1–3) high-quality, non-redundant questions are selected using textual similarity ranking and regex filtering, ensuring contextual coherence and, when needed, visual clarification. Each selected follow-up is then answered by sampling over 10 candidate responses paired with multiple visual options, from which $\mathcal{N}$ (typically 2–4) responses are chosen based on relevance and multimodal quality, with optional user-guided continuations to enhance satisfaction. Repeating this selection process for n rounds yields a tree-structured QA dataset of size $\prod_{i=1}^{n} \mathcal{M}_i \times \mathcal{N}_i$. For more details, see Appendix D.

Quality Control and Pruning. We apply a filtering strategy from multiple perspectives with two key components: the image(-text) Filter, which evaluates each candidate image for visual quality and semantic relevance, and the consistency filter, which preserves content and stylistic coherence across dialogue turns. Finally, we prune the multi-turn paths based on overall quality, coherence, and diversity, yielding a refined set of QA instances for annotation.

Human Annotation. Defining high-quality multi-turn multimodal dialogues is inherently challenging, as it requires assessing response correctness, the coherence of image-text interleaving, and the dynamic nature of human preferences throughout the conversation. We conduct multiple rounds of in-depth discussions with our annotation team regarding existing open-source datasets and prior work on MLLMs. We then identify the following 9 annotation dimensions.

- G1: Context Awareness
- G2: Helpfulenss and Completeness
- G3: Crucial Step Recognition
- G4: Global Image-Text Consistency
- G5: Style Coherence
- L1: Local Image-Text Consistency
- L2: Visual Perceptual Quality
- L3: Contextual Coherence
- L4: Text Quality

Crowdworkers first rate individual turns and then evaluate entire conversations from both local and global perspectives. A **Dual Verification** stage combines dedicated annotator efforts with professional quality control reviews to ensure guideline adherence. Structured **Language Feedback**, which offers concise explanations of scoring rationale, focused critiques, and refinement suggestions, further guides response improvement and substantially enhances annotation reliability.

3 Analysis

Since the INTERMT dataset captures *real* human preferences across multiple dimensions at both *global* and *local* levels, it is meaningful to analyze the correlations among these dimensions, examine the relationship between per-turn preferences and overall evaluation, and further compare human feedback with AI feedback in this section.

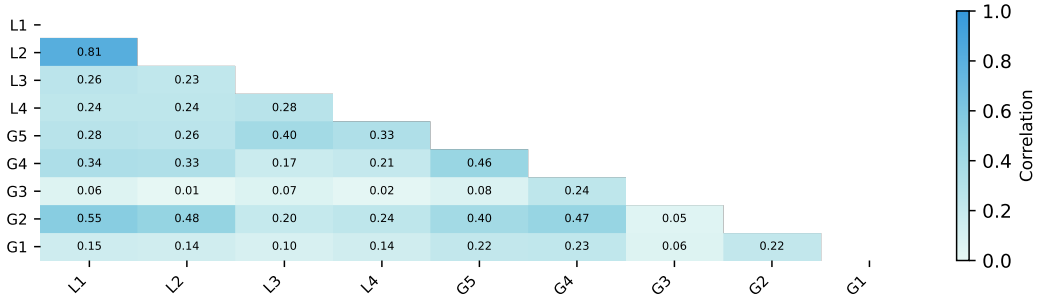

Figure 4: Linear correlation coefficient of different preference annotations.

Correlation Analysis. Figure 4 illustrates the relationship between global and local preference annotation dimensions. We identify three key findings: **(1) Modality perception precedes effective modality fusion**: for both the local-local and local-global correlation, the evaluation of image-text consistency is strongly correlated with visual perceptual quality (up to 0.81). This suggests that before assessing multimodal information, human evaluators tend to prioritize a clear understanding of each individual modality, indicating that a clear perception of individual modalities is a prerequisite for reliable multimodal judgment. **(2) Long-horizon evaluations hinge on coherence and temporal consistency**: for the global-global correlation, metrics such as helpfulness and completeness strongly align with context awareness and global visual consistency, underscoring the importance of maintaining coherent semantics, multimodal information, and consistency with prior conversational context over extended interactions. **(3) Intent grounding drives long-horizon crucial step recognition**: in multi-turn scenarios, models may deviate from the user's core intentions, producing self-directed responses. Despite locally high-scoring and plausible outputs, this leads to stylistic drift and omission of key steps over extended interactions, as demonstrated in the local-global correlation setting.

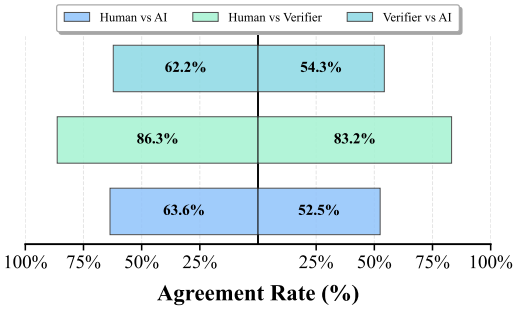

Figure 5: Agreement rates between different roles on local and global levels.

Human Feedback *vs.* AI Feedback. Human-labeled data introduce high cost, which motivates the exploration of MLLMs' potential to assist with evaluation tasks [44]. We develop a pipeline that utilizes advanced API-based models [38, 45, 46, 39, 47, 27]) to produce multidimensional scores from both global and local perspectives. Then, we evaluate the agreement between AI and human

annotators, as well as between AI annotators and expert human verifiers. Agreement scores are then averaged across all pairs for comparative analysis. Experimental results (Fig. 5) show that while AI annotators achieve approximately 60% agreement on local evaluations, their consensus with humans on global (longer-horizon) tasks is markedly lower. This indicates current MLLMs struggle to match human judgments in multi-turn, multimodal scoring. Until AI feedback efficacy is firmly established, replacing human annotation remains inadvisable.

4 Inspiring Future Research

INTERMT lays the groundwork for advancing research on aligning human values in *multi-turn multi-modal understanding and generation* tasks, potentially inspiring new research directions. Building on real human data provided by INTERMT, we identify several promising directions:

- **Modeling long-horizon values.** How can we model long-horizon, interleaved multimodal preferences by leveraging the *local* and *global* human annotations in INTERMT?
- **Aligning dynamic human values**: How can we design algorithms that effectively incorporate real human feedback from INTERMT to assess and enhance the performance of MLLMs?

In this section, we present several baseline approaches that address the above questions, with the goal of fostering further research and demonstrating the utility of our dataset.

4.1 Preference Modeling for Multi-turn Interleaved Multimodal Scenarios

A widely adopted approach for modeling human preferences is to employ a preference predictor grounded in the Bradley–Terry (BT) model [48]. However, when extending to multi-turn settings, new challenges arise—particularly in capturing the dynamics of evolving user preferences across turns. Moreover, traditional outcome-level reward signals often fail to generalize in purely textual domains [49], let alone in complex multimodal settings involving interleaved understanding and generation. INTERMT incorporates both *local* and *global* human annotations in multi-turn, multimodal interactions, leading us to investigate efficient preference modeling methods.

Inspired by [50, 51], we investigate two strategies for modeling long-horizon preferences in multi-turn multimodal scenarios: *prefix preference* and *chain-based preference*. Details of formulations can be seen in Appendix H. Our findings, presented in Figure 6, suggest that modeling fine-grained *local (turn-level)* preferences is more effective in capturing human values and achieving better alignment. In contrast, directly modeling *global (conversation-level)* preferences often fails to reflect these nuanced preferences, especially in complex, long-horizon scenarios.

Local *vs.* Global Preference Transfer. We examine the bidirectional transfer between turn-level (*local*) and conversation-level (*global*) human preferences. As shown in Figure 6, both *local-to-global* and *global-to-local* transfers are effective, since multi-turn questions typically hinge on the seed question's intent. However, *global-to-local* transfer is consistently easier and better aligned with actual preferences. We attribute this to the greater stability of global preferences—reflecting users' overarching tendencies—whereas local preferences are short-term and more context-dependent, making *local-to-global* transfer more challenging.

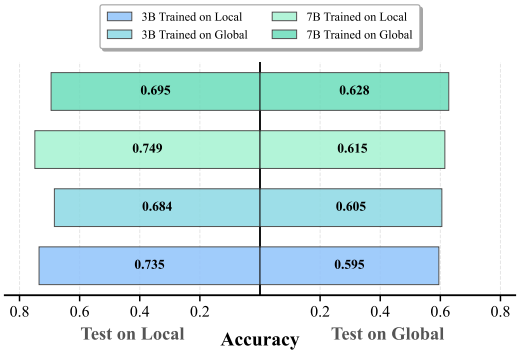

Figure 6: Judge models trained and evaluated on different dataset.

Multi-turn Scaling Law of Turn-based Judge Moderation *Can we accurately capture users' intentions and latent preferences with a limited number of conversational turns, thereby improving the modeling of long-term values?* Such capabilities are crucial for building general-purpose AI assistants, which need to understand and predict users' needs across diverse contexts, adapting to changing preferences over time. We investigate whether the discriminative power of judge models, trained on the first k turns, improves in subsequent turns (from $k+1$ to N) and exhibits *scaling laws*.

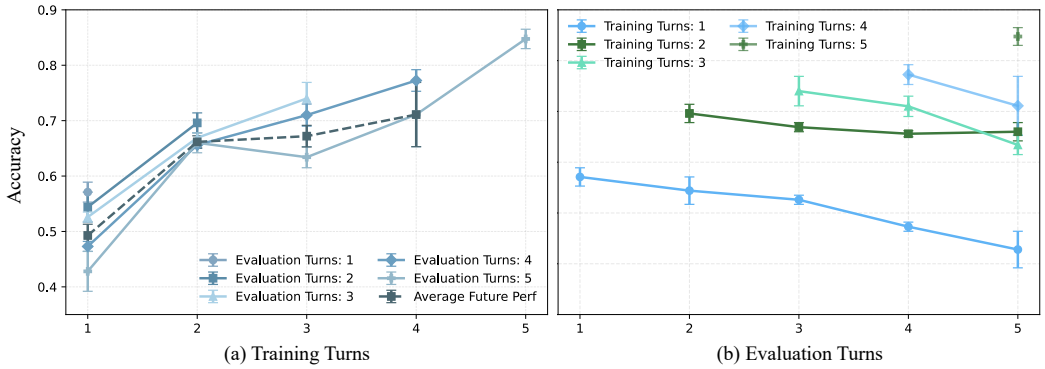

Figure 7: Scaling laws of judge models. As training turns increase, model's ability to predict future preferences improves (left), while generalization diminishes as evaluation turns increases (right).

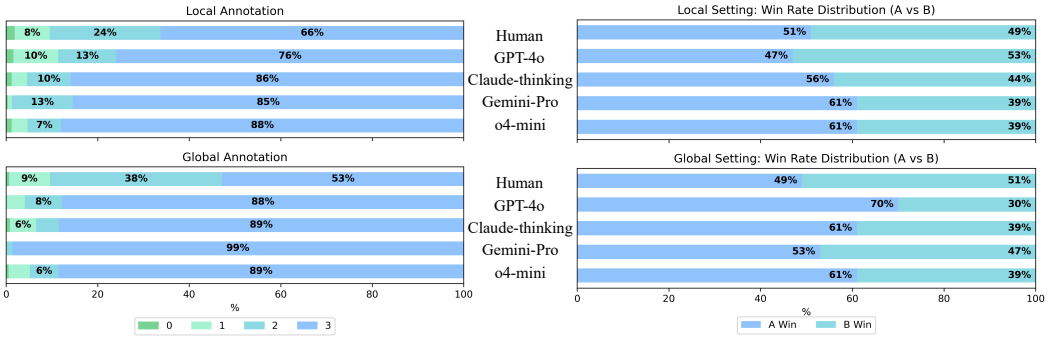

Figure 8: Score and length distribution comparison.

The results reveal two key insights: (1) Multi-turn judge moderation exhibits a generalization effect linked to the number of turns. As shown in Figure 7 (a), for evaluation turn k, as the number of preceding turns increases from 1 to $k-1$, the model's accuracy continues to improve, with average future performance rising, indicating that training on multi-turn data with a limited number of turns can generalize to longer horizons. (2) Regarding the number of turns in the training data, the generalization effect shows a diminishing trend. As demonstrated in Figure 7 (b), training with k turns does improve performance for $k+1 \rightarrow T$ turns, but this effect diminishes as the number of turns increases. The decline is due to three factors: diminishing returns as the model struggles with long-term preferences, contextual drift as earlier turns lose relevance, and the evolving interaction between user intentions and latent preferences.

4.2 MLLM as a Judge and INTERMT-BENCH

Do MLLMs truly understand what is desirable in multi-turn, multimodal interactions and how to align with human values? This task is particularly challenging due to the absence of multimodal benchmarks that capture human preferences in multi-turn settings. Inspired by [52, 44] and leveraging genuine feedback from INTERMT, we introduce INTERMT-BENCH to assess MLLMs' alignment with human values in multi-turn, multimodal tasks. INTERMT-BENCH comprises three distinct tasks: *Scoring Evaluation*, *Pair Comparison*, and *Crucial Step Recognition*. Details can be found in Appendix G.

Results and Takeaways We evaluated 6 advanced MLLMs for their ability to assist in judgment for multi-turn multimodal interactions, considering the nine dimensions proposed above. The results reveal key observations: **Existing models still face challenges in aligning with long-horizon human values, but they perform more accurately in evaluating local, fine-grained preferences**. As shown in Table 1, all models exhibit significant gaps in performance compared to humans in both Score Evaluation and Pair Comparison tasks. However, the models demonstrate better accuracy when assessing local dimensions rather than global dimensions, suggesting that capturing fine-grained (*e.g.*,

Table 1: Overall performance comparison of different MLLMs in three judgment tasks of INTERMT-BENCH. All reported pearson similarity values exhibit a p-value below 0.05, indicating a statistically significance confidence level.

Settings	MLLMs	Local Setting					Global Setting					
		L1	L2	L3	L4	Avg.	G1	G2	G3	G4	G5	Avg.
Scoring Evaluation	Gemini-Flash*	0.346	0.107	0.119	0.173	0.186	0.163	0.042	0.051	**0.246**	0.005	0.101
	Gemini-Flash* (+reason)	**0.361**	0.072	0.122	0.168	0.181	-0.038	0.083	0.139	0.199	0.048	0.086
	GPT-4.1	0.264	0.095	0.242	0.269	0.218	0.215	0.216	0.084	0.044	0.049	0.122
	GPT-4.1 (+reason)	0.281	0.094	0.272	0.271	0.229	0.215	0.255	0.217	0.216	0.050	**0.191**
	GPT-4o	0.291	**0.131**	0.277	0.268	**0.242**	0.254	0.167	0.137	0.139	0.069	0.153
	GPT-4o (+reason)	0.290	0.091	0.252	**0.280**	0.228	0.183	0.243	0.194	0.086	0.072	0.156
	Gemini-Pro*	0.273	0.079	0.258	0.168	0.194	**0.285**	0.240	-0.024	0.235	**0.145**	0.176
	Gemini-Pro* (+reason)	0.274	0.070	0.304	0.211	0.215	0.239	**0.267**	0.195	0.129	0.060	0.178
	Claude-thinking*	0.299	0.044	0.262	0.229	0.209	0.172	0.140	0.175	0.150	0.069	0.141
	Claude-thinking* (+reason)	0.291	0.023	0.254	0.214	0.196	0.207	0.260	0.183	0.155	-0.001	0.161
	o4-mini	0.334	0.062	0.306	0.134	0.209	0.169	0.161	0.120	0.096	0.028	0.115
	o4-mini (+reason)	0.326	0.056	**0.322**	0.151	0.214	0.215	0.229	**0.347**	0.137	0.016	0.189
Pair Comparison	GPT-4.1	0.541	**0.589**	0.508	0.484	0.531	0.540	0.520	0.530	**0.590**	**0.563**	**0.549**
	GPT-4.1 (+reason)	0.550	0.584	0.501	0.521	**0.539**	0.520	0.520	0.477	0.513	0.540	0.514
	GPT-4o	0.513	0.488	0.499	0.510	0.503	0.560	0.517	**0.550**	0.543	0.470	0.528
	GPT-4o (+reason)	0.500	0.537	**0.511**	0.509	0.514	0.542	0.490	0.545	0.522	0.528	0.525
	Gemini-Pro*	0.533	0.521	0.496	0.533	0.521	**0.562**	**0.566**	0.523	0.505	0.505	0.532
	Gemini-Pro* (+reason)	0.526	0.528	0.513	0.514	0.520	0.548	0.562	0.495	0.522	0.538	0.533
	Claude-thinking*	0.561	0.568	0.508	0.502	0.535	0.539	0.523	0.518	0.521	0.528	0.526
	Claude-thinking* (+reason)	**0.567**	0.550	0.506	0.519	0.536	0.512	0.522	0.512	0.547	0.512	0.521
	o4-mini	0.556	0.549	0.508	**0.536**	0.537	0.552	0.498	0.522	0.518	0.495	0.517
	o4-mini (+reason)	0.521	0.564	0.522	0.513	0.530	0.534	0.510	0.507	0.512	0.483	0.509

turn-level) human preferences is crucial for both evaluation and alignment with human dynamic and long-horizon values. However, there is cause for optimism: current MLLMs exhibit near-human-level performance (4.38/5) in recognizing task completion and aligning with human intent (*i.e.*, *Crucial Step Recognition*), providing potential solutions for long-term value alignment.

Induced Bias and Hallucination. Consistent with [44], we identify issues related to bias and hallucination: **Position Bias**, where models consistently favor responses in specific positions (*e.g.*, the first answer), often influenced by training data that places correct answers at the beginning or end of prompts [53], and **High-Score Bias** [44], where models tend to assign higher scores to entire multi-turn communications. These issues, particularly in long-horizon tasks, may hinder the model's ability to capture differences between extended conversations, thereby posing challenges in modeling long-horizon human values and potentially leading to safety concerns [54].

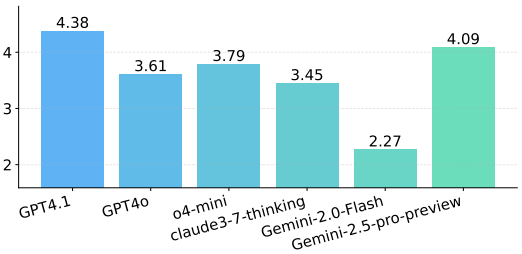

Figure 9: Results of *Crucial Step Recognition*.

5 Conclusion and Outlook

This work introduces INTERMT, the first human preference dataset designed for multi-turn, multimodal understanding and generation tasks, capturing human feedback at both local (turn-level) and global (conversation-level) granularities across nine dimensions. We also present INTERMT-BENCH to evaluate the capability of advanced MLLMs in assisting with judging such complex interactions. We find that modeling fine-grained local (turn-level) preferences is generally more effective in capturing human values and achieving better alignment compared to directly modeling global (conversation-level) preferences. Analyzing preference transfer, we observe that while both local-to-global and global-to-local transfers are effective, global-to-local transfer is consistently easier and better aligned with actual preferences. Another key observation is the multi-turn scaling law of judge moderation: as the number of training turns increases, the model's ability to predict future preferences improves, while its generalization ability diminishes with longer evaluation horizons. Future work is essential to extend INTERMT to encompass these additional modalities, moving closer to a holistic representation of communications dynamics.

References

[1] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[2] Matthew Turk. Multimodal interaction: A review. *Pattern recognition letters*, 36:189–195, 2014.

[3] Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[4] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024.

[5] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. NExt-GPT: Any-to-any multimodal LLM. In *Forty-first International Conference on Machine Learning*, 2024.

[6] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[7] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

[8] Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*, 2023.

[9] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[11] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[13] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

[14] Jiaming Ji, Jiayi Zhou, Hantao Lou, Boyuan Chen, Donghai Hong, Xuyao Wang, Wenqi Chen, Kaile Wang, Rui Pan, Jiahao Li, et al. Align anything: Training all-modality models to follow instructions with language feedback. *arXiv preprint arXiv:2412.15838*, 2024.

[15] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[16] Jiaming Ji, Boyuan Chen, Hantao Lou, Donghai Hong, Borong Zhang, Xuehai Pan, Tianyi Alex Qiu, Juntao Dai, and Yaodong Yang. Aligner: Efficient alignment by learning to correct. *Advances in Neural Information Processing Systems*, 37:90853–90890, 2024.

[17] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

[18] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.

[19] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.

[20] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, 2024.

[21] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 564–572, 2024.

[22] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8228–8238, 2024.

[23] Yi-Fan Zhang, Tao Yu, Haochen Tian, Chaoyou Fu, Peiyan Li, Jianshu Zeng, Wulin Xie, Yang Shi, Huanyu Zhang, Junkang Wu, et al. Mm-rlhf: The next step forward in multimodal llm alignment. *arXiv preprint arXiv:2502.10391*, 2025.

[24] Minqian Liu, Zhiyang Xu, Zihao Lin, Trevor Ashby, Joy Rimchala, Jiaxin Zhang, and Lifu Huang. Holistic evaluation for interleaved text-and-image generation. *arXiv preprint arXiv:2406.14643*, 2024.

[25] Yun He, Di Jin, Chaoqi Wang, Chloe Bi, Karishma Mandyam, Hejia Zhang, Chen Zhu, Ning Li, Tengyu Xu, Hongjiang Lv, et al. Multi-if: Benchmarking llms on multi-turn and multilingual instructions following. *arXiv preprint arXiv:2410.15553*, 2024.

[26] Ved Sirdeshmukh, Kaustubh Deshpande, Johannes Mols, Lifeng Jin, Ed-Yeremai Cardona, Dean Lee, Jeremy Kritz, Willow Primack, Summer Yue, and Chen Xing. Multichallenge: A realistic multi-turn conversation evaluation benchmark challenging to frontier llms. *arXiv preprint arXiv:2501.17399*, 2025.

[27] OpenAI. o4 mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

[28] Herbert Paul Grice. Logic and conversation. *Syntax and semantics*, 3:43–58, 1975.

[29] Barbara Grosz, Aravind Joshi, and Scott Weinstein. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 1995.

[30] Herbert H Clark and Susan E Brennan. Grounding in communication. *psycnet*, 1991.

[31] Constituency Parsing. Speech and language processing. *Power Point Slides*, 2009.

[32] David Rood Traum. *A computational theory of grounding in natural language conversation*. University of Rochester, 1995.

[33] Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2406.10462*, 2024.

[34] Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, Lijuan Wang, and Huaxiu Yao. MMIE: Massive multimodal interleaved comprehension benchmark for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[36] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[37] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

[38] OpenAI. GPT4o. https://openai.com/index/hello-gpt-4o/, 2024.

[39] Google Deepmind. Gemini 2.0 Pro Flash. https://aistudio.google.com/prompts/new_chat?model=gemini-2.0-flash-exp, 2025.

[40] Anthropic. Claude 3. https://www.anthropic.com/news/claude-3-family, 2024.

[41] Black Forest Labs. black-forest-labs/flux (github repository). https://github.com/black-forest-labs/flux, 2024.

[42] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[43] Google Deepmind. Claude 3. https://developers.googleblog.com/en/experiment-with-gemini-20-flash-native-image-generation/, 2025.

[44] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[45] OpenAI. GPT4.1. https://openai.com/index/gpt-4-1/, 2025.

[46] Google Deepmind. Gemini 2.5 Pro. https://deepmind.google/technologies/gemini/pro/, 2025.

[47] Anthropic. Claude 3.7 Sonnet. https://www.anthropic.com/news/claude-3-7-sonnet/, 2025.

[48] Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

[49] Lior Shani, Aviv Rosenberg, Asaf Cassel, Oran Lang, Daniele Calandriello, Avital Zipori, Hila Noga, Orgad Keller, Bilal Piot, Idan Szpektor, et al. Multi-turn reinforcement learning from preference human feedback. *arXiv preprint arXiv:2405.14655*, 2024.

[50] Tianyi Qiu, Fanzhi Zeng, Jiaming Ji, Dong Yan, Kaile Wang, Jiayi Zhou, Yang Han, Josef Dai, Xuehai Pan, and Yaodong Yang. Reward generalization in rlhf: A topological perspective. *arXiv preprint arXiv:2402.10184*, 2024.

[51] Weibin Liao, Xu Chu, and Yasha Wang. TPO: Aligning large language models with multi-branch & multi-step preference trees. In *The Thirteenth International Conference on Learning Representations*, 2025.

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[53] Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*, 2023.

[54] Cem Anil, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. *Advances in Neural Information Processing Systems*, 37:129696–129742, 2024.

[55] Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

[56] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[57] Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. Bot-adversarial dialogue for safe conversational agents. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2950–2968, 2021.

[58] Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36:24678–24704, 2023.

[59] Jiaming Ji, Donghai Hong, Borong Zhang, Boyuan Chen, Josef Dai, Boren Zheng, Tianyi Qiu, Boxun Li, and Yaodong Yang. Pku-saferlhf: A safety alignment preference dataset for llama family models. *arXiv e-prints*, pages arXiv–2406, 2024.

[60] Josef Dai, Tianle Chen, Xuyao Wang, Ziran Yang, Taiye Chen, Jiaming Ji, and Yaodong Yang. Safesora: Towards safety alignment of text2video generation via a human preference dataset. *arXiv preprint arXiv:2406.14477*, 2024.

[61] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663, 2023.

[62] Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, Lingpeng Kong, and Qi Liu. Vlfeedback: A large-scale ai feedback dataset for large vision-language models alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6227–6246, 2024.

[63] Yongting Zhang, Lu Chen, Guodong Zheng, Yifeng Gao, Rui Zheng, Jinlan Fu, Zhenfei Yin, Senjie Jin, Yu Qiao, Xuanjing Huang, et al. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*, 2024.

[64] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.

[65] Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*, 2022.

[66] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[67] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*, 2017.

[68] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*, 2018.

[69] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023.

[70] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[71] Google. C4. https://www.tensorflow.org/datasets/catalog/c4, 2022.

[72] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. Obelics: An open web-scale filtered dataset of interleaved image-text documents, 2023.

[73] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021.

[74] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[75] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[76] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409, 2024.

[77] Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[78] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[79] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024.

[80] Hong Zhang, Zhongjie Duan, Xingjun Wang, Yingda Chen, Yuze Zhao, and Yu Zhang. Nexus-gen: A unified model for image understanding, generation, and editing. *arXiv preprint arXiv:2504.21356*, 2025.

[81] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[82] Ziyu Liu, Tao Chu, Yuhang Zang, Xilin Wei, Xiaoyi Dong, Pan Zhang, Zijian Liang, Yuanjun Xiong, Yu Qiao, Dahua Lin, et al. Mmdu: A multi-turn multi-image dialog understanding benchmark and instruction-tuning dataset for lvlms. *arXiv preprint arXiv:2406.11833*, 2024.

[83] Elliot L Epstein, Kaisheng Yao, Jing Li, Xinyi Bai, and Hamid Palangi. Mmmt-if: A challenging multimodal multi-turn instruction following benchmark. *arXiv preprint arXiv:2409.18216*, 2024.

[84] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

[85] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[86] Minbin Huang, Yanxin Long, Xinchi Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. Dialoggen: Multi-modal interactive dialogue system for multi-turn text-to-image generation. *arXiv preprint arXiv:2403.08857*, 2024.

[87] Meera Hahn, Wenjun Zeng, Nithish Kannen, Rich Galt, Kartikeya Badola, Been Kim, and Zi Wang. Proactive agents for multi-turn text-to-image generation under uncertainty, 2025.

[88] Pengfei Zhou, Xiaopeng Peng, Jiajun Song, Chuanhao Li, Zhaopan Xu, Yue Yang, Ziyao Guo, Hao Zhang, Yuqi Lin, Yefei He, et al. Gate opening: A comprehensive benchmark for judging open-ended interleaved image-text generation. *arXiv preprint arXiv:2411.18499*, 2024.

[89] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194*, 2024.

[90] Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. *arXiv preprint arXiv:2401.16745*, 2024.

[91] Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. *arXiv preprint arXiv:2410.13848*, 2024.

[92] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025.

[93] Yiyang Ma, Xingchao Liu, Xiaokang Chen, Wen Liu, Chengyue Wu, Zhiyu Wu, Zizheng Pan, Zhenda Xie, Haowei Zhang, Xingkai yu, Liang Zhao, Yisong Wang, Jiaying Liu, and Chong Ruan. Janusflow: Harmonizing autoregression and rectified flow for unified multimodal understanding and generation, 2024.

[94] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

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

[96] China: Hourly Minimum Wage by Region 2024. https://www.statista.com/statistics/233886/minimum-wage-per-hour-in-china-by-city-and-province, 2025.

[97] Hui Mao, Ming Cheung, and James She. Deepart: Learning joint representations of visual arts. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1183–1191. ACM, 2017.

[98] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020.

[99] Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*, 2023.

[100] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.

[101] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[102] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

[103] Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. Ask for information retrieval: Part i. background and theory. *Journal of documentation*, 38(2):61–71, 1982.

[104] Pia Borlund. The concept of relevance in ir. *Journal of the American Society for Information Science and Technology*, 54(10):913–925, 2003.

[105] David Ellis. A behavioural approach to information retrieval system design. *Journal of documentation*, 45(3):171–212, 1989.

[106] Carol Collier Kuhlthau. *Seeking Meaning: A Process Approach to Library and Information Services*. Libraries Unlimited, 2 edition, 2004.

[107] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424, 1989.

[108] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[109] Joseph Lee Rodgers and W Alan Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66, 1988.

Appendix

Table of Contents

A Related Work

A.1 QA Dataset with Human-Preference Annotation

Human preference annotations are essential for aligning language models with the 3H objectives: helpfulness, harmlessness, and honesty [55, 15, 8]. These preferences are typically converted into reward signals via the Bradley-Terry model [48], facilitating the use of established RL methods [56] or direct policy optimization toward preferred response distributions [17]. A number of datasets offer question–answer pairs with human preference annotations, ranging from safety-focused datasets [57, 56, 58, 59] to multimodal preference datasets [14, 13, 60, 61, 62, 63]. However, preference data in multi-turn dialogue settings remains underexplored. Existing studies primarily focus on multi-turn response generation [49], rather than improving instruction-following quality across turns, particularly in multimodal contexts. INTERMT fills this gap by introducing a dataset specifically designed for preference-based human annotation dataset in multi-turn, multimodal interactions.

A.2 Interleaved Image-Text Dataset

Training on interleaved image-text web documents has shown superior performance compared to simple image-description pairs, as demonstrated by models such as Flamingo [9], Chameleon [11], and MiniGPT-5 [64]. This improvement is attributed to the richer and more meaningful correlations in interleaved documents, underscoring their importance in developing interleaved generation models. However, the training data used in these studies is not publicly available. Recent efforts have focused on constructing interleaved image-text datasets [65, 66, 67, 68, 33]. For instance, MMC4 [69] extends the text-only C4 dataset [70, 71] by incorporating images into text documents. OBELICS [72] collects large-scale data from web pages. However, both datasets suffer from low image-text coherence and a limited number of images per document [33]. Other datasets focus on image-centered question answering [66, 67, 68]; however, their limited task diversity and reasoning depth reduce their suitability for high-quality visual instruction tuning. CoMM [33], which sources data from websites such as WikiHow, emphasizes visual tutorials. Nevertheless, none of these datasets support multi-turn interactions. To address these limitations, we present INTERMT—a multi-turn image-text interaction dataset encompassing diverse tasks, including visual instruction following, image editing, causal reasoning and so on. INTERMT emphasizes image-text coherence, and logical consistency across dialogue turns, aiming to enhance the general instruction-following capabilities of MLLMs.

A.3 Multi-Turn QA Dataset

Recent studies have concentrated on constructing multi-turn dialogue datasets, typically through human-human interactions, to facilitate the development of more effective chat-based AI assistants. These datasets generally incorporate both vision and text modalities [67, 66, 68, 65]. However, these datasets are restricted to image-text inputs and textual multi-turn outputs, which are often collected via crowdsourcing under narrowly defined tasks. Consequently, the resulting data often contain colloquial expressions, making them suboptimal for enhancing instruction-following capabilities. More importantly, these studies lack a principled methodology for constructing multi-turn preference datasets. To advance any-modality MLLMs, there is still a notable scarcity of high-quality vision-language interactive datasets that incorporate human-annotated preference.

A.4 Modeling of Interleaved Image-Text

The advent of multimodal large language models has markedly advanced tasks involving interleaved text-image understanding and generation. Earlier models like DALL-E [73] and Stable Diffusion [74] showcased impressive capabilities in generating high-quality images from textual descriptions, whereas models such as LLaVA [10, 4] achieve notable breakthroughs in image understanding and reasoning via vision instruction tuning. However, previous research has predominantly focused on unidirectional generation—either text-to-image or image-to-text—without addressing interleaved generation scenarios in which text and images are seamlessly integrated within the same input or output. Recent efforts have begun to close this gap [75, 5, 9, 76, 77, 78, 79, 80]. Flamingo [9] introduced image tokens into the language modeling process, whereas Chameleon proposed a unified architecture embedding both modalities into a shared space for multimodal input and output. Emu [76] utilizes Stable Diffusion [74] as an image decoder, thereby enabling generation from interleaved

image-text inputs. Despite these advances, existing models continue to struggle with multimodal contextual consistency—such as semantic coherence and stylistic alignment between images and text [5, 33, 24, 34]. Furthermore, multi-turn dialogue capabilities—such as contextual coherence, modality-aware content selection, and heuristic question generation—remain underexplored. To address these gaps, we present INTERMT—a human preference dataset specifically designed for multi-turn interleaved text-image understanding and generation.

A.5 AI Alignment and RLHF

Aligning LLMs with human preferences is critical for their safe and effective deployment [8]. Among various approaches, supervised fine-tuning (SFT) and reinforcement learning from human feedback (RLHF) have emerged as standard methods for aligning model behavior with human intent [56, 15, 17]. Recent work has extended this alignment framework beyond language-only settings to multimodal scenarios involving both image and text modalities [13, 23, 18, 14]. Such multimodal alignment necessitates addressing challenges like interleaved image-text inputs and outputs, alongside multi-turn interactions that reflect real-world usage. However, the approach to alignment for the *multi-turn interleaved multimodal understanding and generation* setting still remains an open question.

While highly effective for single-turn instruction following, extending RLHF to multi-turn dialogue introduces significant challenges. These include capturing context-dependent preferences that evolve over the conversation, maintaining long-term coherence and consistency, the increased cost and complexity of collecting high-quality multi-turn preference data, and potential reward hacking where the model optimizes for local turn-level rewards at the expense of overall conversational quality [15].

Some works take an initial step toward multi-turn alignment by leveraging conversation-level human feedback in purely textual multi-turn dialogue, mainly focusing on *how to generate better multi-turn dialogue* [49]. However, improving instruction-following abilities for *multi-modal understanding and generation* in *multi-turn* settings still remains an open challenge.

A.6 Evaluating Multi-turn Multimodal Capabilities

Recent advancements in evaluating multi-turn multimodal capabilities of MLLMs have highlighted the need for benchmarks that reflect real-world conversational complexities. Traditional evaluation datasets often focus on single-turn interactions or unimodal inputs, which do not adequately capture the challenges posed by multi-turn, multimodal dialogues.

To address this gap, several benchmarks have been proposed [81, 82, 83, 84, 85, 86, 87, 88]. For instance, MMDU introduces a comprehensive benchmark designed to evaluate MLLMs' abilities in multi-turn and multi-image conversations [82]. It emphasizes the importance of long-context understanding and the integration of multiple images within a single dialogue, pushing models to handle more realistic and complex interactions. ConvBench introduces a hierarchical evaluation framework that assesses LVLMs across three cognitive levels: perception, reasoning, and creativity [89]. This structure enables a nuanced analysis of model performance in multi-turn dialogues, highlighting specific areas for improvement. Similarly, MMMT-IF presents a challenging benchmark focusing on instruction-following in multimodal, multi-turn dialogues [83]. It introduces the Programmatic Instruction Following (PIF) metric, which assesses a model's ability to follow instructions dispersed across long dialogues, requiring the retrieval and reasoning over instructions spread throughout the context. In the realm of language models, MT-Eval offers a comprehensive benchmark to evaluate multi-turn conversational abilities [90]. By analyzing human-LLM conversations, it categorizes interaction patterns and constructs multi-turn queries to assess models' performance in maintaining context and coherence over multiple turns. Collectively, these benchmarks highlight the importance of developing evaluation methods that reflect the intricacies of multi-turn, multimodal interactions. However, a common limitation among them is the primary focus on understanding capabilities, often neglecting the generation aspect of multimodal interleaved information. This oversight presents challenges in providing per-turn and overall feedback judgments, which are crucial for the comprehensive assessment and improvement of MLLMs.

B Data Examples

We conduct an in-depth comparison of both open-source and API-based models, including Janus [91, 92, 93] and Gemini [94, 95], on multi-turn multimodal understanding and generation tasks (Case Study). We further present representative examples of multi-turn QA and preference-annotated instances in INTERMT (Examples). Please refer to https://pku-intermt.github.io/ for more details.

C Data Details

C.1 Existing Asset Licenses

The INTERMT dataset is released under the **CC BY-NC 4.0** License. Some seed questions used for eliciting multi-turn communications are sourced from open-source datasets, as shown in Table 3, all of which are also under the **CC BY-NC 4.0** License. Additionally, we have obtained data from Wikihow and Ehow through legitimate means. The real images included in our dataset are sourced from Google Images and Pinterest, all of which were acquired legally.

C.2 Data Access

Our homepage is available at https://pku-intermt.github.io/. The dataset consists of three parts hosted on Huggingface:

- INTERMT: A human preference dataset contains 15,604 unique seed questions across diverse categories, 52.6k multi-turn interleaved vision-language QA instances, and 32,459 sets of multi-dimensional human preference annotations. It is available at https://huggingface.co/datasets/PKU-Alignment/InterMT.
- INTERMT-BENCH: A carefully constructed dataset for evaluating MLLMs in assisting judgment capabilities under multi-turn multimodal understanding and generation. It is available at https://github.com/cby-pku/InterMT.
- INTERMT-JUDGE: A tool that leverages INTERMT preference modeling for multi-turn multimodal judge scenarios, achieving a consistency rate of 75%, outperforming most advanced API-based models. It is available at https://huggingface.co/PKU-Alignment/InterMT-Judge.

C.3 Instituional Review Board (IRB)

The human annotations and data usage in this work have received approval from the Institutional Review Board (IRB) of the Institute for Artificial Intelligence at Peking University, and the relevant materials are included in the supplementary files.

Fair and Ethical Labor We employed 30 full-time crowdsourced workers with substantial experience in multimodal annotation for leading commercial language models. To acknowledge their contributions, we adopted a fair and transparent compensation scheme. The estimated average hourly wage ranged from USD 8.56 to USD 10.23 (XE rate as of 2025/05/13), substantially exceeding the local minimum wage of USD 3.66 in Beijing, PRC [96]. In accordance with local labor laws, workers followed a standard Monday-to-Friday schedule, working eight hours per day with weekends off.

Fair Use of Dataset and Identifying Potential Negative Societal Impacts The INTERMT project has undergone a thorough review and audit by the Academic Committee of the Institution for Artificial Intelligence at Peking University. An Institutional Review Board (IRB) has evaluated this work to ensure that the use of the INTERMT dataset adheres to principles of fairness and integrity. During dataset construction, we conducted NSFW filtering to enhance internal safety; however, we acknowledge that absolute safety cannot be guaranteed. Given that multimodal data may pose greater societal risks than pure text data, we believe it is necessary to consider implementing safeguards for sensitive content, such as adopting Hugging Face's gated dataset access settings. We are committed to developing safe and beneficial AI technologies and strongly oppose any misuse that hinders human progress. We unequivocally condemn malicious use of the INTERMT dataset and advocate for its responsible and ethical use.

C.4 Comparison with other datasets

As shown in Table 2, compared to existing multimodal datasets, INTERMT is the first human preference dataset designed for multi-turn multimodal interactions. Each multi-turn QA instance includes interleaved textual and visual content in both inputs and outputs, with an average of 5.33 images per conversation, simulating complex real-world human-AI communication scenarios.

Table 2: Comparison between INTERMT with other image-text datasets. Inter-I: interleaved image-text input; Inter-O: interleaved output; Multi-I: multi-turn for input; Multi-O: multi-turn for output.

Dataset	Data Scale	Inter-I	Inter-O	Multi-I	Multi-O	#Num Categories	Preference
CoMM [33]	227k	Yes	Yes	No	No	5	No
OBELITICS [72]	141M	Yes	Yes	No	No	200	No
MMC4 [69]	101.2M	Yes	Yes	No	No	30	No
Visual Dialogue [66]	120k	Yes	No	Yes	Yes	80	No
IGC [67]	4.2k	Yes	No	Yes	Yes	N/A	No
Image-Chat [68]	202k	Yes	No	Yes	Yes	215	No
MM-Dialogue [65]	1.08M	Yes	Yes	Yes	Yes	4184	No
RLHF-V [13]	5.7k	Yes	No	No	No	-	Yes
INTERMT (**Ours**)	32.4k	Yes	Yes	Yes	Yes	15+	Yes

D Data Collection

In this section, we detail the data construction process of INTERMT, as illustrated in Figure 3. The pipeline consists of four main stages. **Stage I:** Seed questions are collected from open-source corpora, web content, and human-authored sources. These are then filtered based on perceived quality, topical diversity, and potential for multi-turn expansion. **Stage II:** We apply iterative prompting of MLLMs, augmented with external tools (*e.g.*, editing, retrieval, and generation), to produce answer elaborations and follow-up questions, constructing candidate QA trees. **Stage III:** Human annotators perform both per-turn (local) and conversation-level (global) assessments—evaluating dimensions such as quality, coherence, context awareness, and completeness—to prune and select preferred branches. **Stage IV:** The selected branches are reorganized into deep, coherent QA trees (with depth ≥ 5), forming the final multi-turn QA pairs used for model training.

D.1 Prompt Collection

INTERMT is built from 100k image-text QA instances sourced from three primary channels: approximately 72.1% are derived from open-source corpora—namely, publicly available datasets related to vision-language tasks [14, 23, 33] (Table 3 summarizes the open-corpus vision-language datasets used in our pipeline, along with the input-output formats of their original annotations.) ; around 22.8% originate from legally scraped web content (*e.g.*, multimodal platforms such as WikiHow and Pinterest); and the remaining 5.1% are contributed by researcher-curated, human-written prompts. These instances span a wide range of vision-language tasks, including activity generation, data visualization, and table analysis.

Table 3: Collected datasets and their corresponding task types. We select various datasets to ensure that the *seed question*s encompass diverse query styles and originate from a broad range of sources.

I/O Format	TI2T	TI2TI	T2I
Datasets	LLaVA-Instruct-150K [1]	LLaVA-Instruct-150K [1]	DiffusionDB [100]
	ART500K [97]	RLHF-V [13]	MS COCO [101]
	MovieNet [98]	MM-RLHF [23]	HPDv2 [102]
	RLHF-V [13]	Align-Anything-200K [14]	Pick-a-Pic-v2 [61]
	ShareGPT4V [99]	CoMM [33]	Align-Anything-200K [14]

Grounded in theoretical frameworks from linguistics, human-computer interaction, and cognitive psychology [28, 29, 30, 31, 32], we identify five prototypical scenarios that commonly lead to

Table 4: Prototypical scenarios that commonly lead to multi-turn conversations in real-world multimodal contexts, grounded in theories of information retrieval and communication.

Task Type	Concept
Unclear Cognition	Based on Belkin's ASK model, users are in an "anomalous state of knowledge" during retrieval [103]. They recognize knowledge gaps but cannot clearly articulate their needs. Multi-turn dialogue assists in clarifying their goals through guided interaction.
Repeated Attempts Due to Unsatisfactory Answers	According to Borlund's interactive IR model, retrieval is a dynamic, iterative process [104]. Users may re-query after unsatisfactory results. Multi-turn dialogue enables iterative feedback and refinement of information needs.
Complex Tasks Requiring Stepwise Progression	Drawing from Ellis's behavioral model and Kuhlthau's ISP model, complex tasks require phased progress [105, 106]. Multi-turn dialogue supports task decomposition and information integration, helping users build knowledge step by step.
Exploratory or Companion-like Interaction	Bates's Berrypicking model illustrates non-linear, evolving information behavior [107]. Users follow shifting interests rather than fixed goals. Multi-turn dialogue provides contextual guidance and emotional engagement in open-ended exploration.
Cross-modal Multiturn Interaction	This involves integrating language and visual modalities. User needs may be embedded across modalities, requiring multi-turn dialogue to semantically align and interpret multimodal information for accurate understanding and task resolution.

multi-turn conversations in real-world multimodal contexts: (1) incomplete or unclear user cognition; (2) follow-up queries prompted by unsatisfactory initial responses; (3) complex tasks that require incremental, stepwise reasoning; (4) open-ended or companion-like dialogic interactions; and (5) cross-modal mismatches arising from latent inconsistencies between image and text modalities or the need for integrated cross-modal reasoning. Table 4 presents formal definitions of these sceanarios.

Guided by these scenarios, we filter, diversify, and rewrite the original image-text QA instances, resulting in 15,604 unique *seed questions*, which serve as initial prompts for generating iterative, multi-turn conversations. Figure 11 presents the system prompt used with GPT-4o [38] to evaluate the suitability and potential for multi-turn communication, as well as to assist in filtering and rewriting the original data. Figure 10 illustrates the distribution of *seed questions* across more than 15 distinct vision-language tasks. Table 5 provides definitions and representative examples for each task category.

D.2 Iterative Questioning and Response Generation

Iterative Questioning To simulate realistic multi-turn communications, the construction process begins with carefully designed *seed questions* that possess the potential to trigger extended multimodal conversations. In subsequent rounds, agents adopt a *Socratic questioning* strategy, generating context-aware follow-up questions based on the prior conversation history. These follow-ups fall into five common categories frequently observed in real-world multimodal conversations: emotional responses that convey empathy or affective engagement, inquiries that deepen or elaborate on prior content, challenges that test logical consistency or factual accuracy, task decomposition for complex problem solving, and natural terminations when the topic has been sufficiently explored. At each turn, a pool of *10* candidate questions is generated by diverse agents, and a subset of $\mathcal{M}$ (typically 1–3) high-quality and low-redundancy candidates is selected based on textual similarity ranking and regular-expression-based filtering of malformed text. The resulting follow-up questions consistently maintain contextual coherence, ensure conversation continuity, and often leverage visual modalities when necessary to enhance clarity or specificity. Figure 12 presents the system and user prompt for generating follow-up questions.

Response Generation In each turn, every follow-up question ($\mathcal{M}$ per round) is addressed by sampling *10+* candidate responses from diverse agent models. Each response is paired with multiple visual candidates, forming a multimodal answer set. Outputs are expected to be complete, accurate, concise, and helpful, with optional user-guided continuations (*e.g.*, *Would you like a further explanation ?*) to improve user satisfaction. A subset of $\mathcal{N}$ responses (typically 2–4) is selected based on contextual relevance and multimodal quality. Repeating this process across n rounds yields a

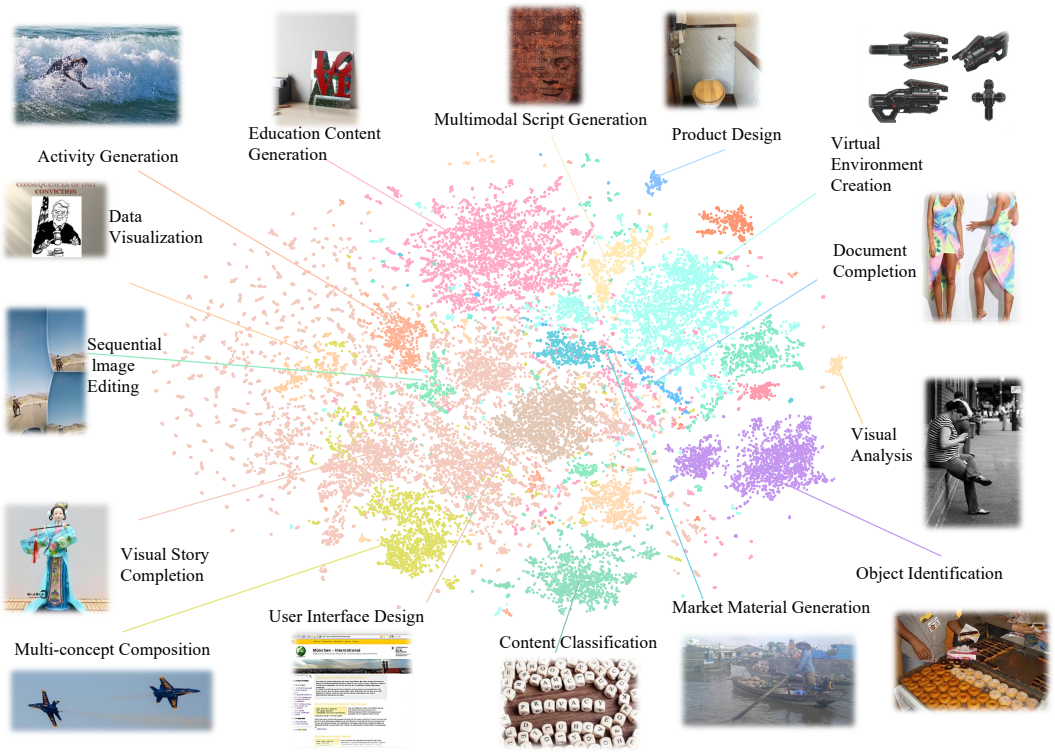

Figure 10: Seed question distribution of INTERMT. Each seed question facilitates **multi-turn** QA, encompassing a broad range of over 15 vision-language **understanding** and **generation** tasks.

tree-structured QA dataset, where each seed question expands into $\prod_{i=1}^{n} \mathcal{M}_i \times \mathcal{N}_i$ multi-turn paths. Figure 13 presents the system and user prompt for generating multimodal responses.

D.3 Agent Construction

The agent workflow is built upon a combination of strong open-source models (*e.g.*, Qwen2-VL [35], Qwen2.5-VL [36], Gemma3 [37], and LLaVA-1.5 [4] series) alongside leading API-based models (*e.g.*, GPT-4o [38], Gemini-2.0-Flash [39] and Claude-3.7-Sonnet-Thinking [40]).

Specifically, the following model list are used to construct agents for iterative question generation and response: API-based models include GPT-4o [38], Gemini 2.0 Flash [39], Claude 3.7 Sonnet (both thinking and standard variants) [40]. Open-source models include Qwen-2-VL-72B-Instruct [35], Qwen2.5-VL-32B-Instruct [36], Gemma3-27B-Instruct [37], and LLaVA-v1.5-7B [10].

To support diverse multimodal operations, three types of image-centric tools are integrated: (1) text-to-image generators (*e.g.*, FLUX.1-Schnell [41] and Stable-Diffusion [42]) for producing high-quality images based on prompts; (2) an image editing API (*e.g.*, Gemini-2.0-flash-exp-image-generation [43]) capable of cropping, highlighting, and modifying images; and (3) web-based retrieval interfaces (*e.g.*, Google Images, Pinterest) for sourcing real-world visuals.

D.4 Quality Control and Pruning

We employ a multi-perspective filtering strategy to ensure the quality and coherence of the dataset, which can be broadly categorized into two types.

- **Image(-Text) Filter**: For single-turn image selection, both visual quality and semantic consistency with the text are critical to ensure the selected image is both legible and contextually appropriate. We adopt an image(-text) filter that integrates visual quality assessment and semantic alignment with the input text to rank and filter the candidate image pool returned by the image tool calling module. Specifically, given candidate images $I = \{i_1, \ldots, i_N\}$ and caption T, we assign each

Table 5: Prompt categories and their definitions in multi-turn interleaved multimodal understanding and generation tasks.

Prompt Category	Definition
Fairstyle Generation	Generate hairstyle designs and styling suggestions based on textual descriptions or reference images. Useful in virtual try-on systems or beauty applications.
Report Generation	Produce structured reports or analytical summaries from multimodal inputs, including images and text. Often used in medical imaging, quality inspection, or news summarization.
Activity Generation	Create interactive activities, games, or engagement schemes tailored to specific topics, scenarios, or user profiles.
Document Completion	Extend or complete existing documents by inferring and preserving their content structure, semantics, and writing style.
Visual Story Completion	Generate coherent continuations or endings for visual narratives based on initial scenes or images.
Multimodal Script Generation	Create instructional or narrative scripts that combine visual and textual components. Common in tutorial videos or AR/VR guides.
Sequential Image Editing	Apply a series of image editing steps in a temporally or logically consistent manner. Suitable for demonstrations or step-wise transformations.
Multi-concept Composition	Integrate multiple concepts, styles, or thematic elements into a unified visual or multimodal output.
Education Content Generation	Generate learning materials, lesson plans, or courseware using multimodal prompts. Can be customized by subject, age group, or learning objectives.
Market Material Generation	Create marketing content such as advertisements, banners, or product showcases leveraging both visual and textual cues.
Content Classification	Organize or categorize multimodal content based on semantic, stylistic, or functional criteria.
Visual Analysis	Analyze visual elements and their interrelations within an image, including object detection, spatial layout, or stylistic attributes.
Creative Writing	Generate creative texts—such as stories, poems, or dialogues—conditioned on visual inputs or scenarios.
Technical Explanation	Provide detailed explanations of technical systems or processes by leveraging both images and text. Often applied in educational or industrial settings.
Product Design	Design new products or optimize existing ones, incorporating visual aesthetics, functionality, and user feedback.
Data Visualization	Translate structured data into visual forms such as charts, diagrams, or infographics to facilitate interpretation.
User Interface Design	Create layouts and elements for user interfaces of digital applications, considering usability and visual coherence.
Virtual Environment Creation	Design and describe immersive virtual spaces or scenes, used in simulation, gaming, or training environments.
Other	User-defined categories not covered above. Allows for flexible extensions based on specific task definitions.

image a score that combines two factors: a rule-based score Rule_j combining multi-dimensions (*e.g.*, resolution, clarity *etc.*), and a semantic coherence score $\widetilde{\text{Coher}}_j$ measuring CLIP-based image–text similarity [108]. The final score is computed as:

$$S(i_j, T) = \alpha\, \text{Rule}_j + (1 - \alpha)\, \widetilde{\text{Coher}}_j. \tag{1}$$

Finally, the image i_{j^*} with the highest score is selected, striking a balance between visual quality and semantic coherence to ensure the image is both visually appealing and contextually appropriate.

System Prompt:
You are a communication analysis agent. Your task is to determine whether a given prompt is likely to trigger a **multi-turn conversation**. Your judgment should be grounded in established discourse and cognitive theories, including Grice's Cooperative Principle, Centering Theory, Cognitive Load Theory, and known issues in multimodal vision-language alignment.
Please answer the following questions for each input prompt:
1. Suitability Judgment: Does the prompt contain characteristics that are likely to elicit follow-up questions, clarification requests, elaboration, or continued user engagement (e.g., due to ambiguity, complexity, or referential uncertainty)? Output either `YES` or `NO`.
2. Rationale (if YES): Briefly explain why this prompt would lead to multi-turn interaction. Your explanation should be based on one or more of the following:

- **Underspecification or Ambiguity:** The prompt lacks sufficient detail or contains vague references, prompting clarification.
- **Cognitive Complexity:** The task is complex enough to require stepwise reasoning or decomposed planning, encouraging follow-ups.
- **Discourse Dynamics:** Topic or referential focus shifts during the interaction, necessitating communication continuity mechanisms.
- **Multimodal Mismatch:** The prompt involves visual and textual inputs whose alignment must be verified interactively.
- **Exploratory Intent:** The prompt expresses a subjective or open-ended goal, inviting elaboration, negotiation, or perspective sharing.

Provide your answer in the following format:
```
Judgment: [YES/NO]
Rationale: [Your explanation here]
```

User Prompt:
Prompt: {...}, Image: <image>, your evaluation:

Figure 11: The prompt for evaluating the suitability and potential of multi-turn communication. This prompt assesses whether an input is likely to elicit multi-turn interactions and provides theoretical justifications grounded in discourse, cognitive, and multimodal communication theories.

- **Consistency Filter**: In multi-turn conversations, consistency with prior turns is crucial, both in content (avoiding contradictions with chat history) and style (*e.g.*, maintaining uniform image aesthetics across turns). Advanced models (*e.g.*, GPT-4o [38] and Gemini-2.0-Flash [39]) are employed to better capture such dependencies and ensure coherent filtering across turns.

We then prune the generated tree-structured multi-turn paths based on overall image quality, sequence coherence, and diversity. Paths that include irrelevant images or excessively divergent follow-up questions are removed, resulting in a refined set of multi-turn QA instances for human annotation.

Rule-based Filtering Given a set of candidate images $I = \{i_1, \ldots, i_N\}$ and an associated text description T, we first extract for each successfully loaded image i_j a collection of raw quality metrics:

$$x_j \in \{Res_j, Clar_j, Bright_j, Cont_j, Color_j\},$$

System Prompt:
You are a large multimodal language model simulating a curious and thoughtful human. You are currently engaged in a conversation with an AI Assistant. You will receive the previous turns of the conversation along with the AI Assistant's latest reply.
Your task is to ask a follow-up question or respond interactively based on the AI Assistant's most recent response. Before asking a question, you should first try to understand the conversation history and the User's intent to help you generate a better question for the User. Then, select one of the following interaction categories that best describes your intent:

- **[Emotional Response]**: express emotions, empathy, encouragement, or reactive questions
- **[Follow-up]**: dig deeper or extend the previous answer
- **[Challenge]**: question the logic, detail, or validity of the answer
- **[Step-by-step Task]**: break down a complex task and guide to the next step
- **[END]**: choose to end the conversation when the topic has been fully explored
- **[Other: XXX]**: define your own category if needed

If helpful, you may reference selected modalities to support your question. Use the following format to include them: `<Modality, brief description>` Examples: `<Image, diagram of a volcano>`, `<Audio, sound of rain>`, `<Video, cat jumping over a box>`
Output Format:
`[[Category]][[Your Question]]`
Examples:
```
[[Follow-up]][[You mentioned that volcanic eruptions are often
preceded by earthquakes.  Can we use seismic data to predict
eruptions in advance?]]
```
If you believe no further question is necessary, conclude the conversation with: `[[END]][[Some words to end the conversation]]`

User Prompt:
```
Chat History:  {chat_history}
Selected Modalities:  {selected_modalities} (default = text,image)
Last Turn Response:  {last_turn_response}
```

Figure 12: System and user prompts for follow-up question generation.

where

$$Res_j = W_j \times H_j, \tag{2}$$

$$Clar_j = \mathrm{Var}\big(\mathrm{Laplacian}(\mathrm{Gray}(D_j))\big), \tag{3}$$

$$Bright_j = \mathrm{Mean}\big(\mathrm{Gray}(D_j)\big), \tag{4}$$

$$Cont_j = \mathrm{Std}\big(\mathrm{Gray}(D_j)\big), \tag{5}$$

$$\begin{aligned} Color_j = {} & \sqrt{\big(\mathrm{Std}(R_j - G_j)\big)^2 + \big(\mathrm{Std}(0.5(R_j + G_j) - B_j)\big)^2} \\ & + \; 0.3\sqrt{\big(\mathrm{Mean}(R_j - G_j)\big)^2 + \big(\mathrm{Mean}(0.5(R_j + G_j) - B_j)\big)^2}. \end{aligned} \tag{6}$$

Min–Max Normalization. Each metric is normalized to $[0, 1]$ via

$$\widetilde{x}_j = \frac{x_j - \min_i x_i}{\max_i x_i - \min_i x_i}.$$

Denote the normalized scores $\widetilde{Res}_j, \widetilde{Clar}_j, \widetilde{Bright}_j, \widetilde{Cont}_j, \widetilde{Color}_j$.

System Prompt:
You are a multimodal AI assistant. Your job is to generate helpful, engaging, and clear responses based on user input, which may include text, images, audio, or video.
Instructions:

- Understand the user's intent by analyzing **all input modalities**.
- Provide a **complete, accurate, concise, and helpful** response.
- **Use multimodal outputs purposefully**, to enhance clarity, immersion, or user experience.
- If the `AllowedModalities` list includes non-text types, incorporate **at least one** of them when relevant.
- Clearly mark non-text content using: `<[Modality], brief description>`
- Examples: `<Image, diagram of a volcano>`, `<Audio, sound of rain>`, `<Video, cat jumping over a box>`
- **Optionally conclude** your response with a natural follow-up question or suggestion to **encourage multi-turn conversation**.
- Besides user question, you will also receive a list of previous user questions and assistant responses (chat history). You should base your response on the chat history.
- You should also consider the user's intent and the chat history when generating your response.

Modality Control:

- Only use modalities listed in `AllowedModalities`.
- If `AllowedModalities = []`, generate a text-only response and briefly explain why no other modality is included.
- Never fabricate modality content or reference unsupported types.

Input may include:

- A text prompt
- Optional and random input modalities for the user prompt (image, audio, video)

Always ground your response in the actual input provided.

User Prompt:

```
User Prompt: {prompt}
Previous User Questions and Assistant Responses: {chat_history}
AllowedModalities: {allowed_modalities} (default = text,image)
```

Figure 13: System and user prompts for multimodal interactive answer generation.

Brightness Penalty. We impose a smooth penalty proportional to deviation from the acceptable brightness range $[30, 220]$:

$$Pen_j = -200 \cdot \Big[\max(0, 30 - Bright_j) + \max(0, Bright_j - 220)\Big],$$

and then normalize

$$\widetilde{Pen}_j = \frac{Pen_j - \min_i Pen_i}{\max_i Pen_i - \min_i Pen_i} \quad \big(\widetilde{Pen}_j \in [0, 1]\big).$$

Rule-Based Quality Score. The normalized quality score is a weighted sum of the normalized metrics plus the penalty. We set fixed weights summing to 1:

$$w_1 = 0.20, \quad w_2 = 0.25, \quad w_3 = 0.15, \quad w_4 = 0.25, \quad w_5 = 0.15, \quad \sum_{k=1}^{5} w_k = 1.$$

Thus the normalized quality score is

$$Rule_j = 0.20\,\widetilde{Res}_j + 0.25\,\widetilde{Clar}_j + 0.15\,\widetilde{Cont}_j + 0.25\,\widetilde{Color}_j + 0.15\,\widetilde{Pen}_j. \tag{7}$$

Text–Image Coherence. We use `openai/clip-vit-base-patch32` to compute a CLIP-based similarity

$$Coher_j = \text{CLIP\_score}(D_j, T), \quad Coher_j \in [-1, 1]$$

Then the raw CLIP similarity $Coher_j \in [-1, 1]$ is shifted and scaled to $[0, 1]$ by

$$\widetilde{Coher}_j = \frac{Coher_j + 1}{2}.$$

Final Selection. Combining normalized quality and coherence, the total score is

$$S(i_j, T) = \alpha\, Rule_j + (1 - \alpha)\,\widetilde{Coher}_j, \quad \alpha \in [0, 1]. \tag{8}$$

We fix the balance parameter $\alpha = 0.7$ (putting 70% weight on visual quality and 30% on semantic coherence). Finally, we select

$$j^* = \arg\max_j S(i_j, T), \quad i_{j^*} \text{ as the output (fallback to } i_1 \text{ if all loads fail).}$$

Thus, each i_j is evaluated both on intrinsic visual quality and on semantic alignment with T, and the maximum-scoring image is selected.

D.5 Human Preference Annotation

Defining high-quality multi-turn multimodal communications is inherently challenging, as it involves evaluating response accuracy, the coherence of image-text interactions, and the evolving nature of human preferences over the course of the conversation. We conduct multiple rounds of in-depth discussions with our annotation team, regarding existing open-source datasets and prior work on MLLMs. We then identify three key criteria: (1) *image-text coherence and helpfulness* — responses should align well with visual content and be logically complete; (2) *contextual consistency* — each turn should maintain thematic relevance, preserve core topics, and ensure stylistic continuity; (3) *long-horizon evaluation* — both local (turn-level) and global (conversation-level) quality should be assessed to evaluate each turn's contribution to overall conversation.

- G1: Context Awareness
- G2: Helpfulenss and Completeness
- G3: Crucial Step Recognition
- G4: Global Image-Text Consistency
- G5: Style Coherence
- L1: Local Image-Text Consistency
- L2: Visual Perceptual Quality
- L3: Contextual Coherence
- L4: Text Quality

Based on these principles[2], we evaluate multi-turn QA instances from both local and global perspectives. Crowdworkers first rate single turns across four sub-dimensions, then assess the full conversation across five dimensions, finally providing preference labels based on aggregated scores.

[2] G_i is used for global evaluation, and L_i is for local evaluation.

Dual Verification All annotations are first completed by a dedicated full-time annotation team and subsequently reviewed by a professional quality control unit, which collaborates closely with our researchers to ensure guideline adherence. Additionally, our team manually audits 20% of the data. Although the task involves inherently subjective human judgments, this dual verification stage primarily aims to suppress annotation noise and improve data quality. Appendix E presents the annotation documents.

Table 6: **Human agreement across different sub-dimensions.**

	G1	G2	G3	G4	G5	L1	L2	L3	L4
w/o Language Feedback (%)	82.1 ± 2.0	81.4 ± 2.8	83.7 ± 3.5	80.6 ± 2.3	83.2 ± 2.5	82.8 ± 3.1	84.1 ± 2.7	85.9 ± 2.6	86.3 ± 3.0
w/ Language Feedback (%)	–	–	**88.3 ± 1.3**	–	–	**87.2 ± 1.5**	**85.8 ± 1.6**	–	–

One More Thing: Language Feedback Building on prior works [60, 14, 16], we incorporate human-written natural language feedback into the annotation process. Each feedback instance includes: (i) *reason*, explaining the rationale behind the assigned score; (ii) *critique*, identifying strengths and weaknesses of the response based on detailed evaluation criteria; and (iii) *refinement*, offering suggestions for improving the image or textual quality of the response. The structured feedback protocol significantly improves inter-annotator agreement by 5–10 percentage points, as shown in Table 6.

E Annotation Documents

E.1 Withdraw

What is an incorrect answer?

- Providing a link that cannot be accessed.
- Giving the current date, but it does not match the actual date.
- Providing a highly time-sensitive response, while the actual situation has changed. For example, listing the "top ten trending songs" without specifying the inability to access the most recent data will be considered invalid.
- Factual errors that contradict objective reality.

What is an invalid question? We carefully examine invalid questions during data validation and continuously update the definition of invalid questions.

- Incomplete questions, such as containing only a single word like "I" or "Hello."
- Questions that lack context, making them difficult to understand.
- Requests for analysis of a given text or context without actually providing the text or context.
- Questions that contain factual errors, rendering the question itself invalid.

What is an ungradable question?

- Highly subjective tasks, such as creative writing, where there is no objective standard for determining quality.
- Questions that exceed the annotator's knowledge level, such as those involving advanced coding, finance, computer science, or physical laws.
- Two questions with answers that are too similar, such as "apple" and "apple." (only differing by punctuation).

What are questions that require web search? Many questions require searching the internet, especially when objective facts are needed.

E.2 Features of Annotated Data

Core Characteristics

- Multi-turn Dialogue: The dataset includes both short dialogues (2-3 turns) and long dialogues (5 or more turns) to accommodate various task requirements.
- Image-Text Interaction: The data includes a combination of text input, image input, and mixed text-image input.
- Scoring System: The scoring follows a fine-grained scale, where each dialogue turn is independently scored, and the final composite score reflects overall performance.

Data Types

- Image-Text Input with Multi-turn Text Output (e.g., step-by-step optimization of a design image).
- Text Input with Multi-turn Image Output (e.g., describing an object and generating images from various perspectives).
- Image-Text Input with Multi-turn Image-Text Output (e.g., design modification process with visual outputs).
- Image Input with Multi-turn Text Output (e.g., providing detailed interpretation or analysis of an uploaded image).

E.3 Annotation Guidelines

E.3.1 Overall Response Evaluation

Context Awareness **Definition**: The model should retain and understand the dialogue history to ensure contextual coherence, rather than treating each turn as an isolated interaction.

Examples:

- In *visual storytelling* tasks, the model should maintain consistent characters, settings, and plot lines.
- In *design revision* tasks, the model should remember the user's previous requests and avoid repeating suggestions that were previously rejected.

Scoring Criteria:

- **0 points**: The model completely ignores the context; responses are irrelevant or contradict the dialogue history.
- **1 point**: The model partially recalls context but exhibits noticeable information loss or inconsistency in roles.
- **2 points**: Context is mostly preserved, with occasional minor inconsistencies.
- **3 points**: Full understanding of context; responses are logically coherent, with no information loss or contradictions.

Helpfulness and Completeness **Definition**: Measures how well the model's textual and visual outputs follow task instructions and provide complete information to fulfill the user's request. This also includes the logical structure of the response. In multi-turn image-text interactions, the model should accurately follow all instructions and ultimately deliver a complete solution.

Example:

- *Task*: Cake design improvement
  - *User*: Please help me improve this design (uploads image)
  - *AI*: Suggests adding frosting (but no new image generated) $\rightarrow$ Deduct points
  - *User*: Please show me the modified 3D rendering

  - *AI*: [Generates an image of the cake with frosting] → Full score

Scoring Criteria:

- **0 points**: The model fails to meet the user's needs; responses are irrelevant or severely incorrect, making the task unachievable.
- **1 point**: Partially satisfies the user's request but lacks critical content or contains major errors that hinder task completion.
- **2 points**: Largely completes the task but has minor omissions or inaccuracies that affect the final outcome.
- **3 points**: Fully and accurately fulfills all user requirements; information is comprehensive, logically structured, and free from errors or omissions.

Crucial Step Recognition **Definition**: In multi-turn interactions, the model must accurately identify and complete crucial steps, avoiding irrelevant or incorrect information.

Example:

- *Task*: Step-by-step guidance for drawing a cat
  - *Crucial steps*: Sketch outline → Refine facial features → Adjust proportions → Apply color
  - *Incorrect*: Model asks the user to color before the outline is drawn
  - *Correct*: Model guides the user through steps in a logical order

Scoring Criteria:

- **0 points**: Key steps are entirely incorrect or omitted, preventing task completion.
- **1 point**: Some steps are inaccurate, though the task may still proceed with effort.
- **2 points**: Overall step sequence is reasonable, with minor deviations or logical flaws.
- **3 points**: All crucial steps are correctly identified and ordered, with no redundancy or omissions.

Global Image-Text Consistency **Definition**: In multi-turn image-text interactions, textual descriptions should align closely with the generated images. Inconsistencies between text and images, or failing to generate images when required, result in lower scores.

Example:

- *Task*: AI-generated interior design plan
  - *User*: Please provide a modern-style living room design
  - *AI*: [Generates image, but the style does not match] → Deduct points
  - *User*: Please change the sofa color to dark grey
  - *AI*: [Generates image with dark grey sofa] → Full score

Scoring Criteria:

- **0 points**: Images are completely unrelated to the text, or necessary images are missing.
- **1 point**: Partial relevance, but with significant mismatches (e.g., incorrect color or structure).
- **2 points**: Largely consistent, with minor deviations.
- **3 points**: Perfect alignment between text and images, with no inconsistencies.

Style Coherence **Definition**: Assesses the consistency of style and subject representation across generated images, including texture, color harmony, lighting, rendering style, physical properties, clothing, and behavior. It penalizes visual repetition, such as overly similar outputs or duplicated elements within a single image. In multi-turn interactions, generated images should exhibit stylistic coherence across turns, with smooth transitions and no abrupt changes.

Special Case:

- If only one turn includes an image while the other does not, visual style coherence is **not** affected. In such cases, assign a **default score of 3 points**.

Scoring Criteria:

- **–1 point**: The task required image generation, but none was provided.
- **0 points**: Images exhibit entirely different styles, tones, rendering, or subject traits, resulting in visual dissonance.
- **1 point**: Some stylistic or subject consistency, but with clear discrepancies (e.g., sudden tone changes, mismatched rendering, or inconsistent subject traits).
- **2 points**: Style, tone, and subject representation are generally consistent, with minor variations that do not affect overall coherence.
- **3 points**: All images are highly consistent in style, tone, quality, and subject representation; visual transitions are smooth and contextually appropriate.

E.3.2 Turn-level Evaluation Metrics

Local Image-Text Consistency **Definition**: In a single dialogue turn, the textual description should closely match the generated image(s), ensuring the text accurately reflects the visual content without ambiguity or misleading information.

Applicable Scenarios:

- If the turn includes multiple images, evaluate the overall consistency of the text with all images. Individual image feedback can be added as needed (e.g., [3,1] Image 1: accurate; Image 2: inconsistent).
- If no image is generated, evaluate based on task requirements:
  - If image generation was expected but omitted, assess the inconsistency between the text and the missing visual content.
  - If the task (e.g., Visual Analysis) does not require image generation, assess consistency between the input image and the text.
- Otherwise, default evaluation compares the answer text with the image(s) generated in that turn.

Scoring Criteria:

- **0 points**:
  - Text is irrelevant to the image(s) or contains major factual errors;
  - Key descriptions are missing or completely incorrect (e.g., referencing nonexistent objects or scenes);
  - Text may cause significant misunderstanding.
- **1 point**:
  - Text is partially related to the image(s), but includes clear errors or misleading descriptions;
  - Covers part of the image content but omits or misrepresents key details or relationships;
  - Reader must infer or adjust understanding to align with the image(s).
- **2 points**:
  - Text generally matches the image(s), with minor local inaccuracies (e.g., imprecise attribute descriptions or slight omissions);
  - Does not hinder overall comprehension, but lacks precision upon close inspection.
- **3 points**:
  - Text is highly aligned with the image(s), covering all key elements and details;
  - Free from factual errors or ambiguity; the description is natural and coherent.

Visual Perceptual Quality **Definition**: Evaluates the visual realism, naturalness, and absence of distortion or artifacts in the generated image(s). Focuses on whether the image structure, colors, and composition realistically simulate the physical world, avoiding unnatural artifacts.

Applicable Scenarios:

- In multi-image outputs, assign a unified score for overall quality. If image quality varies significantly, provide per-image feedback as needed (e.g., [3,1] Image 1: good; Image 2: distorted).
- If no image is generated, assess any image provided in the user prompt. If the image has issues, point them out in the textual answer.

Scoring Criteria:

- **0 points**:
  - Obvious artifacts (e.g., disconnections, misalignments), severe distortions (e.g., highly unrealistic shapes), or structural errors (e.g., unbalanced proportions, illogical composition);
  - Unnatural color rendering (e.g., harsh color blocks, abnormal tones);
  - Lighting does not follow physical laws, severely affecting image recognizability.
- **1 point**:
  - Image is mostly recognizable but contains localized severe flaws;
  - Examples: anatomical errors (e.g., limb dislocation), inconsistent local color (e.g., banding, strong noise), or small rendering failures;
  - Overall naturalness is compromised, affecting visual coherence.
- **2 points**:
  - Image is generally natural and coherent; structure, color, and lighting are mostly reasonable;
  - Minor local imperfections such as rough edges, small artifacts, or slight blurring that do not affect overall perceptual quality.
- **3 points**:
  - Image is visually realistic and natural;
  - Well-structured, smooth color transitions, physically consistent lighting;
  - No visible artifacts, distortions, or flaws; overall aesthetics and details are excellent.

Text Quality **Definition**: Measures the clarity, coherence, and correctness of the output text. Includes grammar, spelling, readability, consistency with instructions and context, and absence of redundancy. Responses should be logically sound, well-structured, and clearly expressed, avoiding abrupt transitions or repetition.

Scoring Criteria:

- **0 points**: Text is disorganized, lacks logic, and is hard to understand; may contain numerous grammar or spelling errors or repetitive content.
- **1 point**: Some parts are logically clear, but the text includes noticeable jumps, omissions, or contradictions that hurt overall readability; may include frequent language errors or redundant expressions.
- **2 points**: The overall logic is reasonable and the flow mostly smooth, but there are minor incoherences; some sentences may require optimization to improve readability.
- **3 points**: Text is logically rigorous, clearly expressed, well-organized, and naturally structured; no obvious jumps or repetition; grammar and spelling are correct, providing a good reading experience.

Contextual Coherence **Definition**: Assesses whether the response in this turn logically continues the dialogue history and remains consistent with prior content, avoiding contradictions.

Scoring Criteria:

- **0 points**: Completely irrelevant or logically inconsistent with previous context.
- **1 point**: Partially relevant but includes clear inconsistencies.
- **2 points**: Mostly coherent, with minor deviations.
- **3 points**: Fully consistent with prior dialogue; no contradictions.

F More details of Annotation

F.1 Annotation Platform

The annotation platform of INTERMT is similar to our sister projects, PKU-SafeRLHF and Beaver-Tails. Based on the specific annotation requirements, we have made appropriate adjustments to the platform, such as adding support for multimodal dialogue inputs across multiple rounds and enabling scoring and preference ranking for each round of communications, as shown in Figure 14. After human annotations, we provide dual verification from both human experts and our researchers.

On the annotation platform, we have provided a comprehensive handbook that includes detailed documentation for the annotation process, as shown in Appendix E, along with summaries and explanations for contentious annotation cases. A withdrawal button is available at the top-right corner of the interface to filter out invalid or meaningless annotation pairs.

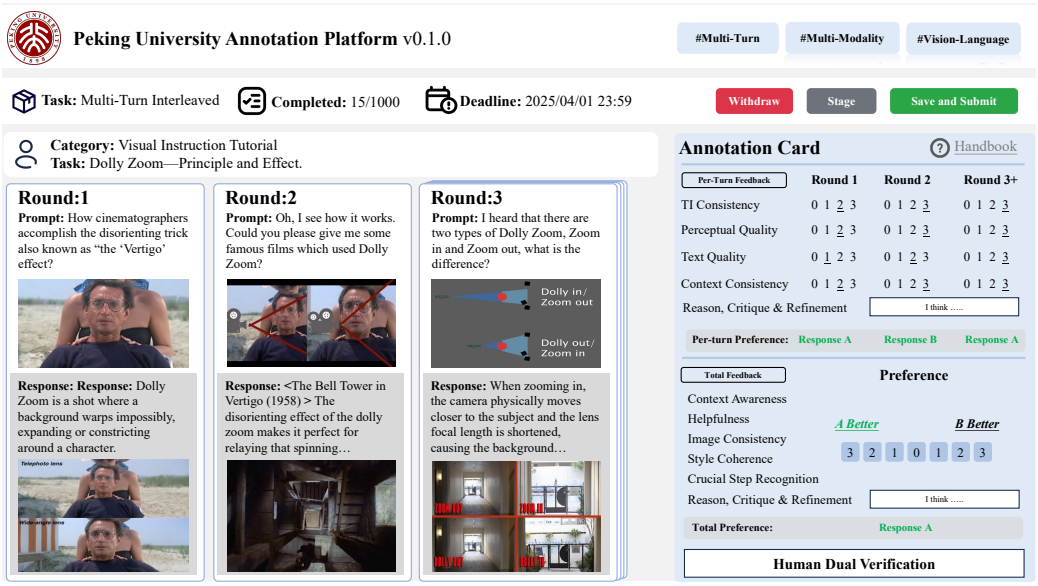

Figure 14: The WebUI of the annotation platform.

Overall, the annotation process consists of three stages:

- **Stage I:** Annotators carefully read and learn the annotation guidelines. They first score each round of a multimodal dialogue (fine-grained scoring) and then determine which round is better between two dialogues (preference ranking).
- **Stage II:** Subsequently, based on the individual round annotations, annotators perform fine-grained scoring for the complete multi-turn multimodal communications and rank which of the two communications is better.
- **Stage III:** The annotation results undergo dual verification by human experts. Annotations with a low consistency rate are rejected for re-annotation. Qualified annotations are then reviewed by researchers through sampling and auditing. Finally, the human preference dataset is finalized.

F.2 Details on Data Labeling Services

Building upon the successes of previous projects such as BeaverTails [58], SafeSora [60], PKU-SafeRLHF [59], and Aligner [16], we again collaborated with the professional annotation service provider AIJet Data. We did not directly interact with the crowdworkers; instead, AIJet managed the entire annotation process. Capitalizing on their expertise in annotating textual data, AIJet curated a dedicated team of experienced annotators tailored to the needs of our project. In light of the task's complexity, we established a contract with a rate above the industry average to prioritize the engagement of qualified personnel. To ensure consistent annotation quality, we supplied AIJet with comprehensive guidelines aimed at standardizing and refining the labeling criteria.

G More Details about INTERMT-BENCH

G.1 Review of Human Annotation Dimensions

We first revisit the key dimensions of **multi-turn** interleaved multimodal **understanding** and **generation**, which also serve as the annotation criteria for our human-labeled dataset. The evaluation in INTERMT-BENCH is conducted with respect to these dimensions, guided by genuine human feedback. Specifically, G_i denotes global evaluation, while L_i corresponds to local evaluation.

- G1: Context Awareness
- G2: Helpfulenss and Completeness
- G3: Crucial Step Recognition
- G4: Global Image-Text Consistency
- G5: Style Coherence
- L1: Local Image-Text Consistency
- L2: Visual Perceptual Quality
- L3: Contextual Coherence
- L4: Text Quality

G.2 Judge Settings and Metrics

The dataset includes multi-turn multimodal interleaved communication histories and human-annotated ground truth. Evaluated models must assess the conversation at both the turn and conversation levels across nine dimensions, following a set of guidelines. *Scoring Evaluation* requires the model to assign scores on a 0-3 scale, with evaluation based on agreement and Pearson similarity [109, 52, 44]. *Pair Comparison* directly compares two individual turns or entire conversations, without considering ties, and is evaluated for accuracy against human judgments. *Crucial Step Recognition* addresses a key challenge in multi-turn conversations: accurately identifying the user's intent and determining whether it has been fulfilled, evaluated by the score provided by judge according to the human-annotated reference answers.

Note We use the following notation conventions to refer to proprietary models evaluated in our experiments: *Gemini-Flash** refers to `Gemini-2.0-Flash`, *Gemini-Pro** denotes `Gemini-2.5-Pro-preview`, and *Claude-thinking** corresponds to the `Claude-3.7-Sonnet (thinking)` model.

G.3 MLLM as a Judge

Inspired by [44, 52], we leverage genuine human-annotated data collected in INTERMT to construct INTERMT-BENCH, a benchmark designed to evaluate the alignment between models and human values in multi-turn multimodal interaction scenarios. Our evaluation focuses on three key aspects: *Score Evaluation*, *Pair Comparison*, and *Crucial Step Recognition*. The system and user prompts used for *Score Evaluation* and *Pair Comparison* are illustrated in Figure 15, 17 and Figure 19,21, respectively.

We also examine the effect of prompting models to generate rationales on scoring accuracy (Figure 16, 18 and Figure 20, 22. For *Score Evaluation*, we quantify the alignment between model-assigned and human-assigned scores using the Pearson correlation coefficient, computed as follows:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{9}$$

where x_i and y_i denote the scores assigned by the model and human annotators, respectively, and $\bar{x}$ and $\bar{y}$ are their corresponding means.

System Prompt:
You are a scoring model for evaluating the overall quality in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please score the conversation history based on the scoring criteria.
Your output needs to be:

[Evaluation Criterion$_1$, score1], [Evaluation Criterion$_2$, score2], ...

Example:
Evaluation list:
[context_awareness, helpfulness, crucial_step_recognition, global_image_text_consistency, style_coherence]
Output:
[context_awareness, score1], [helpfulness, score2], [crucial_step_recognition, score3], [global_image_text_consistency, score4], [style_coherence, score5]
<Annotation Documents>

User Prompt:
Now, please evaluate the conversation history based on the scoring criteria. And output the result in the format of:

[Evaluation Criterion$_1$, score1], [Evaluation Criterion$_2$, score2], ...

Figure 15: System and user prompts for global evaluation in multi-turn multimodal communications (score only).

Table 7: Human agreement percentage with different judge models. Each judgment is independently reviewed by different annotators.

Settings	MLLMs	Local Setting					Global Setting					
		L1	L2	L3	L4	Avg.	G1	G2	G3	G4	G5	Avg.
Score Evaluation	Gemini-Flash*	0.430	0.625	0.783	0.827	0.666	0.702	0.573	0.593	0.089	0.665	0.524
	Gemini-Flash* (+reason)	**0.437**	**0.626**	0.783	**0.828**	**0.669**	0.702	0.573	0.621	**0.302**	0.669	**0.573**
	GPT-4.1	0.392	**0.626**	0.785	0.791	0.649	0.685	0.585	0.625	0.069	0.681	0.529
	GPT-4.1 (+reason)	0.401	0.626	0.787	0.786	0.650	0.706	**0.597**	0.613	0.060	0.681	0.531
	GPT-4o	0.400	0.558	**0.791**	0.807	0.639	**0.710**	0.577	0.625	0.052	0.681	0.529
	GPT-4o (+reason)	0.404	0.545	**0.791**	0.812	0.638	0.706	0.585	**0.629**	0.056	0.681	0.531
	Gemini-Pro*	0.401	0.588	0.777	0.705	0.618	0.664	0.587	0.555	0.150	0.660	0.523
	Gemini-Pro* (+reason)	0.408	0.598	0.783	0.705	0.623	0.709	0.559	0.623	0.105	0.636	0.526
	Claude-thinking*	0.406	0.614	0.738	0.674	0.608	0.686	0.556	0.619	0.180	**0.686**	0.546
	Claude-thinking* (+reason)	0.412	0.612	0.736	0.662	0.606	0.682	0.569	0.623	0.205	0.682	0.552
	o4-mini	0.429	0.621	0.774	0.714	0.634	0.627	0.525	0.598	0.108	0.672	0.506
	o4-mini (+reason)	0.428	**0.626**	0.781	0.715	0.638	0.675	0.549	**0.638**	0.128	0.675	0.533

G.4 Details of *Crucial Step Recognition* Evaluation

Crucial Step Recognition evaluates whether a model can accurately identify the user's underlying intent in multi-turn multimodal interactions—typically signaled by the initial seed question—and effectively track the evolving user needs and preferences throughout the dialogue. Moreover, it assesses whether the model can recognize which specific step fulfills the user's core intention. This capability is critical for enhancing task completion and user experience in human-AI interactions.

During evaluation, we collect human-annotated rationales for crucial step recognition as reference answers. Given recent findings that advanced models can achieve human-comparable performance in pairwise response comparison [52, 44], we employ GPT-4o [38] as the judge to perform partial order comparisons between model outputs (Figure 23 presents the system and user prompts for evaluated models) and the reference answers (Figure 24 presents the system and user prompts for

System Prompt:
You are a scoring model for evaluating the overall quality in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please score the conversation history based on the scoring criteria and provide a reason for your score.
Your output needs to be:

$$[\text{Evaluation Criterion}_1,\ \text{Reason},\ \boxed{\text{score1}}],\ [\text{Evaluation Criterion}_2,\ \text{Reason},\ \boxed{\text{score2}}],\ \ldots$$

Example:
Evaluation list:
[context_awareness, helpfulness, crucial_step_recognition, global_image_text_consistency, style_coherence]
Output:

$$[[\text{context\_awareness},\ \text{Reason},\ \boxed{\text{score1}}],\ [\text{helpfulness},\ \text{Reason},\ \boxed{\text{score2}}],\ \ldots]$$

<Annotation Documents>

User Prompt:
Now, please evaluate the conversation history based on the scoring criteria. And output the result in the format of:

$$[\text{Evaluation Criterion}_1,\ \text{Reason},\ \boxed{\text{score1}}],\ [\text{Evaluation Criterion}_2,\ \text{Reason},\ \boxed{\text{score2}}],\ \ldots$$

Figure 16: System and user prompts for global evaluation in multi-turn multimodal communications (with reason).

System Prompt:
You are a scoring model for evaluating the quality of a single turn in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please score the conversation history based on the scoring criteria. Your output must follow this exact format:
Evaluation list: [local_image_text_consistency, visual_perceptual_quality, text_quality, context_coherence]
Output:
[[local_image_text_consistency, $\boxed{\text{score}}$]], [[visual_perceptual_quality, $\boxed{\text{score}}$]], [[text_quality, $\boxed{\text{score}}$]], [[context_coherence, $\boxed{\text{score}}$]]
Where `score` is your numerical rating (0–3).
<Annotation Documents>

User Prompt:
Now, please evaluate this turn based on the scoring criteria. Your score should be between 0 and 3. And output the result in the format:

$$[[\text{Evaluation Criterion}_1,\ \boxed{\text{score1}}],\ [\text{Evaluation Criterion}_2,\ \boxed{\text{score2}}],\ \ldots]$$

Figure 17: System and user prompts for local single-turn evaluation (score only).

System Prompt:
You are a scoring model for evaluating the quality of a single turn in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please score the conversation history based on the scoring criteria. Your output must follow this exact format:
Evaluation list:
[local_image_text_consistency, visual_perceptual_quality, text_quality, context_coherence]
Output:
[[local_image_text_consistency, reason, $\boxed{\text{score}}$]], [[visual_perceptual_quality, reason, $\boxed{\text{score}}$]], [[text_quality, reason, $\boxed{\text{score}}$]], [[context_coherence, reason, $\boxed{\text{score}}$]]
Where `score` is your numerical rating (0–3) and `reason` is your brief justification.
<Annotation Documents>

User Prompt:
Now, please evaluate this turn based on the scoring criteria. Your score should be between 0 and 3. And output the result in the format:

[[Evaluation Criterion$_1$, Reason, $\boxed{\text{score1}}$], [Evaluation Criterion$_2$, Reason, $\boxed{\text{score2}}$], . . .]

Figure 18: System and user prompts for local single-turn evaluation (score with reason).

System Prompt:
You are a judge model for evaluating the overall quality in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please compare the two responses (ResponseA and ResponseB) and give your final preference. Your output needs to follow the format:
[Evaluation Criterion$_1$, $\boxed{\text{ResponseA}}$], [Evaluation Criterion$_2$, $\boxed{\text{ResponseB}}$], . . .
<Annotation Documents>

User Prompt: Now, please evaluate the conversation history based on the scoring criteria. And output the result in the format:
[Evaluation Criterion$_1$, $\boxed{\text{ResponseA}}$], [Evaluation Criterion$_2$, $\boxed{\text{ResponseB}}$], . . . ,

Figure 19: System and user prompts for global comparison evaluation (preference only).

System Prompt:
You are a scoring model for evaluating the overall quality in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please score the conversation history based on the scoring criteria and provide a reason for your score. Your output needs to follow the format:

[Evaluation Criterion$_1$, Reason, $\boxed{\text{ResponseA}}$], . . . ,

<Annotation Documents>

User Prompt:
Now, please evaluate the conversation history based on the scoring criteria. And output the result in the format:

[Evaluation Criterion$_1$, Your Judge Reason, $\boxed{\text{ResponseA}}$], . . . ,

Figure 20: System and user prompts for global comparison evaluation (with reason).

System Prompt:
You are a judge model for evaluating the quality of a single turn in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please compare the two responses (ResponseA and ResponseB) and give your final preference. Your output must follow this exact format:
Evaluation list: [local_image_text_consistency, perceptual_quality, text_quality, contextual_coherence]

$$[\text{local\_image\_text\_consistency}, \boxed{\text{ResponseA}}], \ldots,$$

Where "preference" is your preference between ResponseA and ResponseB.
<Annotation Documents>

User Prompt:
Now, please evaluate this turn based on the scoring criteria. Your preference should be between 0 and 3. And output the result in the format:

$$[\text{Evaluation Criterion}_1, \boxed{\text{ResponseA}}], \ldots, [\text{total\_preference}, \boxed{\text{ResponseA}}]$$

Figure 21: System and user prompts for local turn evaluation without reasoning.

System Prompt:
You are a judge model for evaluating the quality of a single turn in multi-turn visual dialogues. You will receive a conversation history, please read it carefully. Next, I will provide you with an evaluation list and corresponding scoring criteria. Please compare the two responses (ResponseA and ResponseB) and give your final preference. Your output must follow this exact format:
Evaluation list: [local_image_text_consistency, perceptual_quality, text_quality, contextual_coherence]

$$[\text{local\_image\_text\_consistency}, \text{reason}, \boxed{\text{ResponseA}}], \ldots,$$

Where "preference" is your preference between ResponseA and ResponseB and "reason" is your brief justification.
<Annotation Documents>

User Prompt:
Now, please evaluate this turn based on the scoring criteria. Your preference should be between 0 and 3. And output the result in the format:

$$[\text{Evaluation Criterion}_1, \text{Reason}, \boxed{\text{ResponseA}}] \ldots, [\text{total\_preference}, \boxed{\text{ResponseA}}]$$

Figure 22: System and user prompts for local turn evaluation with reasoning.

judge models). Each judgment is subsequently reviewed by three human experts, and only those achieving a predefined agreement threshold are considered valid. Final scores are computed by averaging across all evaluation points.

System Prompt:
You are a crucial step recognition model. You will receive a multi-turn dialogue. Based on the dialogue content, determine which steps are crucial and which are optional. Evaluate the model's performance in recognizing key steps and whether it completed the user's initial task.

Crucial Step Recognition:
Definition: In multi-turn interactions, the model must accurately identify and complete crucial steps, avoiding irrelevant or incorrect information.

Example:

- **Task:** Step-by-step guidance for drawing a cat
- **Crucial steps:** Sketch outline → Refine facial features → Adjust proportions → Apply color
- **Incorrect:** Model asks user to color before the outline is drawn
- **Correct:** Model guides user through steps in a logical order

User Prompt:
You are a crucial step recognition model. You will receive a multi-turn dialogue. Based on the dialogue content, determine which steps are crucial and which are optional. Evaluate the model's performance in recognizing key steps and whether it completed the user's initial task. Crucial Step Recognition **Definition**: In multi-turn interactions, the model must accurately identify and complete crucial steps, avoiding irrelevant or incorrect information. **Example**:
* **Task**: Step-by-step guidance for drawing a cat
* **Crucial steps**: Sketch outline → Refine facial features → Adjust proportions → Apply color
* **Incorrect**: Model asks the user to color before the outline is drawn
* **Correct**: Model guides the user through steps in a logical order

Figure 23: System and user prompts for crucial step recognition in multi-turn dialogue.

G.5 More Results

A little knowledge is a dangerous thing Table 7 reports the human agreement accuracy for *Score Evaluation*. Notably, although models often assign identical scores to those given by human annotators, the resulting Pearson correlation coefficients are relatively low. This suggests that models may be guessing scores rather than capturing the nuanced distinctions in human ratings.

Reasoning Ability Is Not a Panacea. We compare weak reasoning (*i.e.*, providing plausible explanations for the evaluation results) with strong reasoning (*i.e.*, using advanced reasoning models like o4-mini) on the scoring evaluation and pair comparison settings. However, the results are suboptimal. The models' reasoning processes are primarily based on a step-by-step comparison against predefined guidelines, rather than actively identifying potential flaws in the responses. This approach, which differs from the more granular feedback humans provide, leads to misalignments with human judgment.

Divide and Conquer is beneficial for Crucial Step Recognition We observe that models with high scores in *Crucial Step Recognition* tend to adopt a divide-and-conquer approach, meaning they first assess each turn in a multi-turn, multimodal dialogue for problem-solving and alignment with human intent, and then provide an overall conclusion; in contrast, models with lower scores often give more generalized responses.

System Prompt:
You are a **Judge Model** designed to evaluate a model's performance in identifying key steps within multi-turn dialogues. Your task is to compare two inputs: 1. **Reference Answer**: The ideal, ground truth response from a model that accurately represents the correct interpretation of the dialogue. 2. **Model Inference**: A model-generated response to the same multi-turn dialogue, which may differ from or match the **Reference Answer**.

Scoring Criteria:

- **Score Range**: 1 to 5 (where 1 is the lowest, 5 is the highest).
- **How to Score**:
  - **5**: Model Inference is flawless or better than Reference Answer. All key steps correct.
  - **4**: Mostly correct with minor issues.
  - **3**: Partially correct with significant omissions or errors.
  - **2**: Many missing or wrong steps.
  - **1**: Fundamentally incorrect or misinterprets dialogue.

Evaluation Guidelines:

- Focus on key steps driving the dialogue.
- Evaluate clarity, accuracy, and logical flow.
- Determine if the Model Inference aligns with intended meaning.

Additional Notes:

- Note if Model Inference is better or worse than Reference Answer.
- Justify the score with detailed rationale.
- Provide recommendations or point out overlooked steps if necessary.

User Prompt:
Now evaluate the following response and give your score and reason. Your score should be in the range of 1 to 5 and in the format of 'score: [[score]], reason: [[reason]]'. {reference_answer} {model_inference}"

Figure 24: System and user prompts for evaluation and crucial step recognition.

H Experiment Details

All training was conducted on 8 NVIDIA H800 GPUs. We used Qwen2.5-VL-3B-Instruct and Qwen2.5-VL-7B-Instruct as the backbone models for training the judge model. Table 8 presents the key training hyperparameters used in our experiments.

H.1 Preliminaries of Preference Modeling

A widely adopted approach for modeling human preferences is to employ a preference predictor grounded in the Bradley–Terry (BT) model [48]. Given a pair of answers $(\boldsymbol{y}_1, \boldsymbol{y}_2)$ generated from an question $\boldsymbol{x}$, BT model indicates that the human preference distribution p^* [15, 17] can be expressed based on the underlying human reward function $r^*(\boldsymbol{y}, \boldsymbol{x})$ as

$$p^*(\boldsymbol{y}_1 \succ \boldsymbol{y}_2 | \boldsymbol{x}) = \frac{\exp(r^*(\boldsymbol{y}_1, \boldsymbol{x}))}{\exp(r^*(\boldsymbol{y}_1, \boldsymbol{x})) + \exp(r^*(\boldsymbol{y}_2, \boldsymbol{x}))},$$

where $\boldsymbol{x} = (\boldsymbol{x}_\text{I}, \boldsymbol{x}_\text{T})$ and $\boldsymbol{y} = (\boldsymbol{y}_\text{I}, \boldsymbol{y}_\text{T})$ represent the multimodal (image-text) input and output, respectively. Hence, given a human image-text preference dataset $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{y}_w^{(i)}, \boldsymbol{y}_l^{(i)})\}_{i=1}^N$, the training objective for a multimodal reward model $r_\phi(\boldsymbol{y}, \boldsymbol{x})$ parameterized by ϕ is defined as:

$$\mathcal{L}(\phi, \mathcal{D}) = -\mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}_w, \boldsymbol{y}_l) \sim \mathcal{D}} [\log \sigma(r_\phi(\boldsymbol{y}_w, \boldsymbol{x}) - r_\phi(\boldsymbol{y}_l, \boldsymbol{x}))]$$

Table 8: Key training hyperparameters used in our experiments.

Parameter	Value
Number of GPUs	8 × NVIDIA H800
Epochs	3
Batch size (train/eval)	8 / 8
Gradient accumulation steps	1
Gradient checkpointing	True
Learning rate	3e-5
LR scheduler	constant_with_warmup
Warmup ratio	0.03
Adam betas	(0.9, 0.95)
Weight decay	0.0
Mixed precision	bf16=True, fp16=False
Evaluation strategy	epoch
Regularization coefficient	0.001
Freeze vision tower	True
Freeze language model	False
Freeze MM projection layer	False
Max token length	8192

However, when extending to multi-turn settings, new challenges arise—particularly in capturing the dynamics of evolving user preferences across turns. Moreover, traditional outcome-level reward signals often fail to generalize in purely textual domains [49], let alone in complex multimodal settings involving interleaved understanding and generation. INTERMT incorporates both *local* and *global* human annotations in multi-turn, multimodal interactions, leading us to investigate efficient preference modeling methods under this more realistic and challenging scenario.

H.2 Long Horizon Human Value Preference Modeling

Inspired by [50, 51], we investigate two strategies for modeling long-horizon preferences in multi-turn multimodal scenarios: *prefix preference* and *chain-based preference*. Let $\mathcal{D}_{\text{multi-turn}} = \{(\boldsymbol{x}_1^{(i)}, \boldsymbol{y}_1^{(i)}, \cdots, \boldsymbol{x}_{k_i}^{(i)}, \boldsymbol{y}_{k_i}^{(i)})\}_{i=1}^{N}$ denote the multi-turn human image-text dataset, where k_i denotes the number of turns for each conversation. The *prefix-preference* approach models preferences at the *turn level*. Given a prefix of the conversation history, it aims to identify the preferred candidate response for the current turn, effectively capturing fine-grained local preferences. The training objective for the *prefix-preference* reward model $r_{\phi_{\text{prefix}}}(\boldsymbol{y}, \boldsymbol{x})$ is

$$\mathcal{L}(\phi_{\text{prefix}}, \mathcal{D}_{\text{prefix}}) = -\mathbb{E}_{(\boldsymbol{z}, \boldsymbol{y}_k^w, \boldsymbol{y}_k^l) \sim \mathcal{D}_{\text{prefix}}} [\log \sigma(r_{\phi_{\text{prefix}}}(\boldsymbol{y}_k^w, \boldsymbol{z}) - r_{\phi_{\text{prefix}}}(\boldsymbol{y}_k^l, \boldsymbol{z})],$$

where $\boldsymbol{z} = (\boldsymbol{x}_1, \boldsymbol{y}_1, \cdots, \boldsymbol{x}_k)$ stands for the shared prefix of the different conversations, and the *prefix-preference* dataset is denoted as $\mathcal{D}_{\text{prefix}} = \{(\boldsymbol{z}, \boldsymbol{y}_k^w, \boldsymbol{y}_k^l) | (\boldsymbol{z}, \boldsymbol{y}_k^w), (\boldsymbol{z}, \boldsymbol{y}_k^l) \sim \mathcal{D}_{\text{multi-turn}}\}$.

In contrast, the *chain-based preference* approach models preferences at the *conversation level* by comparing complete conversation trajectories conditioned on the same *seed question* $\boldsymbol{x}_1$. It seeks to capture the human's overall intent and preference across the entire multi-turn dialogue. The training objective for the *chain-based preference* reward model $r_{\phi_{\text{chain}}}(\boldsymbol{y}, \boldsymbol{x})$ is defined as,

$$\mathcal{L}(\phi_{\text{chain}}, \mathcal{D}_{\text{chain}}) = -\mathbb{E}_{(\boldsymbol{x}_1, \boldsymbol{w}^w, \boldsymbol{w}^l) \sim \mathcal{D}_{\text{chain}}} [\log \sigma(r_{\phi_{\text{chain}}}(\boldsymbol{w}^w, \boldsymbol{x}_1) - r_{\phi_{\text{chain}}}(\boldsymbol{w}^l, \boldsymbol{x}_1))],$$

where $\boldsymbol{w} = (\boldsymbol{y}_0, \boldsymbol{x}_1, \boldsymbol{y}_1, \cdots, \boldsymbol{x}_k, \boldsymbol{y}_k)$ represents the whole conversation chain and the *chain-based* preference dataset is $\mathcal{D}_{\text{chain}} = \{(\boldsymbol{x}_0, \boldsymbol{w}^w, \boldsymbol{w}^l) | \boldsymbol{y}_0^w \neq \boldsymbol{y}_0^l \bigwedge (\boldsymbol{x}_0, \boldsymbol{w}^w), (\boldsymbol{x}_0, \boldsymbol{w}^l) \sim \mathcal{D}_{\text{multi-turn}}\}$.

H.3 Evaluation Details

Due to the absence of publicly available human-annotated test sets for multi-turn multimodal interactions, we adopt a random 9:1 train-test split strategy, ensuring that no *seed question*s appears in both sets. To investigate the multi-turn scaling law of judge models, we ensure that the compared groups with different numbers of communication turns are matched in both data volume and computational cost. We repeat experiments at varying data scales to validate the robustness of our conclusions.