FRABENCH AND UFEVAL: UNIFIED FINE-GRAINED EVALUATION WITH TASK AND ASPECT GENERALIZATION

Shibo Hong[1,2], **Jiahao Ying**[3], **Haiyuan Liang**[1], **Mengdi Zhang**[2]
Jun Kuang[2], **Jiazheng Zhang**[1], **Yixin Cao**[1*]
[1]College of Computer Science and Artificial Intelligence, Fudan University, [2]Meituan
[3]School of Computer Science, Singapore Management University

ABSTRACT

Evaluating open-ended outputs of Multimodal Large Language Models has become a bottleneck as model capabilities, task diversity, and modality rapidly expand. Existing "MLLM-as-a-Judge" evaluators, though promising, remain constrained to specific tasks and aspects (i.e., specific evaluation criteria such as fluency for text and image quality for images). In this paper, we argue that, on one hand, based on the interconnected nature of criteria, learning specific aspects can generalize to unseen aspects; on the other hand, jointly learning to assess multiple visual criteria and tasks may foster a synergistic effect. To this end, we propose UFEval, the first unified fine-grained evaluator with task and aspect generalization for four evaluation tasks — Natural Language Generation, Image Understanding, Image Generation, and Interleaved Text-and-Image Generation. However, training such a unified evaluator is hindered by the lack of a large-scale, multi-modal, and aspect-level resource. To address this gap, we introduce FRABench, a comprehensive fine-grained evaluation dataset. Specifically, (1) We first construct a hierarchical aspect taxonomy encompassing 112 distinct aspects across the aforementioned four tasks. (2) Based on this taxonomy, we create FRABench, comprising 60.4k pairwise samples with 325k evaluation labels obtained from a combination of human and GPT-4o annotations. (3) Finally, leveraging FRABench, we develop UFEval, a unified fine-grained evaluator. Experiments show that learning on specific aspects enables UFEval to generalize to unseen aspects, and joint learning to assess diverse visual tasks and aspects can lead to substantial mutual benefits.

1 INTRODUCTION

As Multimodal Large Language Models (MLLMs) have shown amazing abilities in human-like question answering (Grattafiori et al., 2024), assessing the quality of their free-form outputs has become increasingly challenging. Automated evaluators, a.k.a. "MLLM-as-a-Judge" paradigm (Saha et al., 2025) have therefore received much research attention (Li et al., 2024c). Despite their progress, we posit two concerns: (1) Current evaluators (Liu et al., 2023b; Ke et al., 2023) are typically tailored to specific aspects, which limits their adaptability to unseen aspects. (2) They are also limited to specific tasks and modalities, which sharply constrains their scope of application. Table 1 shows the detailed comparison with existing evaluators. We intuitively argue that, on one hand, evaluation aspects are inherently interconnected (Fu et al., 2023). Specifically, aspects such as engagement, naturalness, and creativity are closely linked. Thus, similar semantics and evaluation standards can be transferred across diverse tasks. On the other hand, jointly learning to assess multiple visual aspects and tasks may foster synergistic effects (Wang et al., 2025b). For instance, learning object alignment in image captioning improves character consistency evaluation in multi-image scenarios, while progress in image understanding enhances image generation evaluation through better judgments of content quality and contextual appropriateness. This cross-aspect and cross-task synergy motivates the development of a unified, fine-grained evaluator for improved generalization and performance.

*Corresponding author

Table 1: Comparison of our UFEval with recent evaluators. NLG, IU, IG, and ITIG represent Natural Language Generation, Image Understanding, Image Generation, and Interleaved Text-and-Image Generation, respectively. The number of aspects shown for each evaluator is comparable in granularity, with the scope covering their respective target domains. "–" indicates that the number of supported aspects is not explicitly specified.

Method	Task				Modalty		Aspect	Generalizable
	NLG	IU	IG	ITIG	Text	Image		
AUTO-J (Li et al., 2023b)	✓	✗	✗	✗	✓	✗	332	✗
X-Eval (Liu et al., 2023b)	✓	✗	✗	✗	✓	✗	27	✓
Prometheus 2 (Kim et al., 2024a)	✓	✗	✗	✗	✓	✗	-	✗
Themis (Hu et al., 2024b)	✓	✗	✗	✗	✓	✗	50	✓
ImageReward (Xu et al., 2023a)	✗	✗	✓	✗	✗	✓	3	✗
VisionReward (Xu et al., 2024)	✗	✗	✓	✗	✗	✓	37	✗
LLaVA-Critic (Xiong et al., 2024)	✗	✓	✗	✗	✓	✗	–	✗
UFEval (ours)	✓	✓	✓	✓	✓	✓	112	✓

To this end, we propose UFEval, the first unified fine-grained evaluator with task and aspect generalization for four evaluation tasks (i.e., NLG, IU, IG, and ITIG) across 28 sub-tasks. However, training such a unified evaluator requires large-scale, multi-modal, aspect-level evaluation datasets, which are currently unavailable. Therefore, we develop the Fine-grained Aspect Benchmark (FRABench) to address this gap. Specifically, we first conduct a survey on the four tasks to identify key evaluation aspects. We then manually organize, extend, and redefine 112 distinct aspects with hierarchical relations as a universal evaluation taxonomy—aspect tree. Using this aspect tree, we construct FRABench with comprehensive aspect coverage. The aspect tree guides us in selecting multiple relevant aspects for each of the 60.4k pairwise responses, for which we obtain evaluation labels through a hybrid approach combining available human annotations with GPT-4o-assisted completions, ensuring cost efficiency. In total, this process yields 325k evaluation labels. Using FRABench, we develop UFEval, a unified and fine-grained evaluator.

Our experiments show that UFEval exhibits excellent evaluation quality and aspect generalization capabilities. This is attributed to learning multiple aspects and tasks jointly, which yields significant mutual enhancement, as well as to fully leveraging the interconnections between learning and unlearning aspects. By conducting ablation experiments across diverse aspects and task baselines, we observe a gradual improvement in evaluation quality, thereby validating our hypothesis. Additionally, we demonstrate the value of using UFEval for preference data generation to align the outputs of models with human preferences via direct preference optimization (DPO). These results validate the effectiveness of the FRABench as a valuable resource for training unified evaluators.

In summary, our contributions are as follows: (1) We construct FRABench, a large-scale multi-modal aspect-level evaluation dataset to train and test evaluators. (2) Upon FRABench, we develop UFEval, the first unified fine-grained evaluator for multiple tasks assessment with task and aspect generalization. (3) Our experiments show that inter-aspect correlations enable generalizable capability, and learning to assess multiple visual tasks and aspects jointly leads to a synergistic improvement in evaluation performance, while also demonstrating UFEval's value for preference alignment.

2 Related Work

2.1 Coarse-grained Specific-task Evaluation

Previously, coarse-grained evaluators, which produce overall judgments based on one or several aspects, have been widely explored across various evaluation tasks (Zhu et al., 2023; Xu et al., 2023b; Ye et al., 2023; Liu et al., 2024b; Wang et al., 2023a; Jiang et al., 2023). For instance, in NLG, Wang et al. (2023b) first introduces PandaLM, a fine-tuned LLM designed to evaluate pairwise texts based on several aspects. Similarly, Li et al. (2023b) proposes Auto-J, an evaluator capable of handling a broader range of tasks and aspects, supporting both pointwise and pairwise evaluation settings. In image generation evaluation, Xu et al. (2023a) develops ImageReward, an evaluator trained on a large-scale human annotation dataset. Using alignment and fidelity as reference dimensions, ImageReward provides an overall assessment. Despite showing good overall performance, coarse-

grained evaluations lack the granularity needed to diagnose specific model deficiencies and often introduce aspect bias into the evaluation process.

2.2 Fine-grained Specific-task Evaluation

To address the limitations of coarse-grained evaluation, recent studies have shifted toward fine-grained evaluation (Ye et al., 2023; Kim et al., 2024b; Ke et al., 2023; Kim et al., 2023; Li et al., 2023d; Ying et al., 2024; Bai et al., 2023), where MLLMs are fine-tuned on multi-aspect datasets to produce aspect-specific judgments. For example, in NLG, Hu et al. (2024b) proposes Themis, an LLM trained on the GPT-4 annotated NLG-Eval corpus. In IU, LLaVA-Critic (Xiong et al., 2024) is the first fine-grained evaluator integrating diverse criteria, showing strong correlation with GPT-4o. For IG, Xu et al. (2024) introduces VisionReward, a VQA-based evaluator that assesses image quality across fine-grained aspects. While these methods improve over coarse-grained evaluation by targeting specific aspects, they struggle with scalability across different tasks and aspects. Liu et al. (2023b) further investigates aspect generalization through X-Eval, a two-stage learning framework that incorporates auxiliary aspects and demonstrates generalization in NLG. However, its reliance on predefined reference aspects and lack of open-source implementation restricts its reproducibility.

3 Methodology

To develop UFEval, we need to construct a large-scale, multi-modal, and aspect-level evaluation dataset. However, existing datasets predominantly focus on overall quality assessment rather than fine-grained aspect evaluation, limiting the development of such evaluators. To address this gap, we construct FRABench through two main steps: (1) Evaluation Aspect Construction, and (2) Fine-grained Evaluation Dataset Construction. The following sections detail these two steps.

3.1 Evaluation Aspect Construction

3.1.1 Aspect Collection and Extension.

To fully leverage existing aspects, we first collect 28 sub-tasks under the four types of tasks (i.e., NLG, IU, IG, and ITIG). These sub-tasks span all six combinations of input types (text and text-with-image) and output types (text, image, and text-with-image), ensuring comprehensive coverage across multimodal tasks. Subsequently, we collect literature related to each sub-task to gather relevant aspects and, where available, their definitions. Moreover, we define aspects without available definitions according to their practical meaning.

Beyond collecting existing aspects, we extend aspects for the ITIG, which lacks sufficient aspects. To this end, we apply a cross-task transfer by identifying analogous sub-tasks from other categories and adapting their aspects. For example, both story generation (NLG) and visual story completion (ITIG) involve narrative creation, enabling aspects like engagingness to be adapted [1]. Ultimately, we obtain 112 different aspects, with detailed information and sources listed in Appendix A.

3.1.2 Aspect Taxonomy Construction.

To facilitate aspect selection within the dataset construction and evaluation pipeline, we organize the collected aspects into an aspect tree serving as a standardized taxonomy. We first use the "overall" aspect as the root node of the aspect tree, and then divide the remaining aspects into two subtrees based on their task independence and task specificity:

- Universal Aspects (UAs): Aspects in this subtree exhibit task independence, as they primarily focus on assessing the quality of the model's output. They are often modality-specific (for example, fluency for text or fidelity for image) and serve as fundamental quality aspects across tasks.
- Task-specific Aspects (TAs): Aspects in this subtree exhibit task dependence, as they primarily focus on task completion rather than output quality. They are typically closely tied to the task type. For example, engagingness for story generation or accuracy for mathematical reasoning.

[1] Aspects sharing the same term but differing in definitions are sequentially marked with superscripts: † and * .

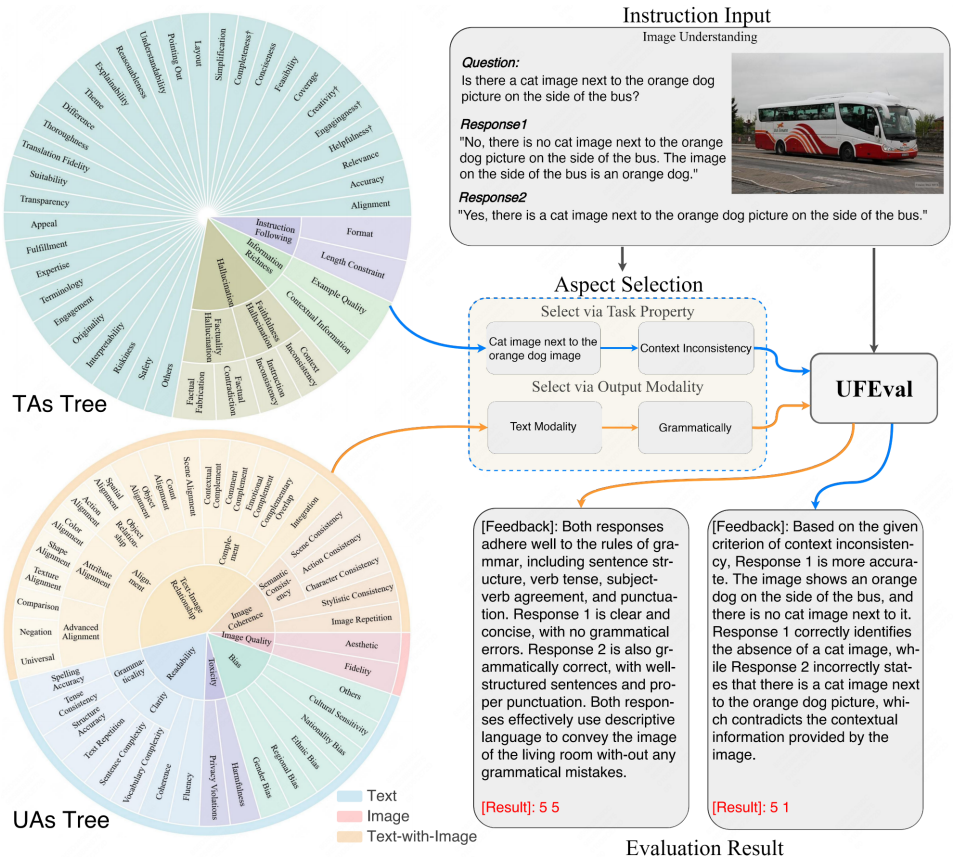
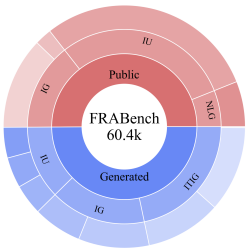


Figure 1: An illustration of the evaluation pipeline. The pipeline consists of two steps: (1) **Aspect Selection**: First, appropriate aspects are selected from the TAs and UAs Trees based on task property and output modality. As illustrated, the question focuses on "cat image next to the orange dog picture", while the image shows no cat adjacent to the orange dog. This discrepancy enables evaluation of the model's ability to avoid hallucinations, specifically its performance on Context Inconsistency. Second, given the text output modality, aspects are selected from the Text Branch in the UAs Tree. (2) **Evaluating**: UFEval generates feedback and scores based on the input content and selected aspects.

For hierarchical construction within the two subtrees, we prioritized aspect trees from peer-reviewed literature to ensure consistency and robustness. Specifically, we adopted the following structures: Readability (Hu et al., 2024a), Bias (DeAlcala et al., 2023), Instruction Following (Zeng et al., 2023), Hallucination (Huang et al., 2025), Alignment (Huang et al., 2023), Complement (Vempala & Preoţiuc-Pietro, 2019), Image Coherence (Liu et al., 2024a), Image Quality (Zhang et al., 2025), Semantic Consistency (Chen et al., 2019) and Toxicity (Gehman et al., 2020).

For the remaining aspects without an established hierarchical structure, we handle them as follows:

- Bidirectional Matching Strategy: First, we check whether the remaining aspect's name appears within the definition of the existing tree's root node. If a match is found, we recursively traverse the child nodes to identify the most specific insertion point and append the aspect as a child of the last matched node. Conversely, if no downward match is found, we verify whether the root node's name appears within the definition of the remaining aspect. In this case, the aspect implies a broader concept and is inserted as the parent of the current root. This applies to the following aspects: Clarity in Readability, Semantic Consistency in Image Coherence, and Integration, Complement, and Alignment in the Text-Image Relationship.
- Creation of new root nodes: For aspects that do not match any existing tree, we established them as independent root nodes to minimize subjective bias and avoid misleading categorizations. These nodes correspond to all green peer-level nodes within the TAs tree in Figure 1.



	Tasks	Dataset	Sub-Tasks	Split	Size
Public	NLG	Auto-J	Summarization, Creative Writing, Rewriting, etc. Question Generation, Title Generation, etc.	Train / FRA-ID FRA-OOD	2.5k / 0.6k 0.6k
	IU	VLFeedback	Image Captioning, Text-rich Understanding, etc. Medical VQA, Academic VQA, etc.	Train / FRA-ID FRA-OOD	14.5k / 2.4k 1.2k
	IG	GenAI-Bench ImageRewardDB	Text-to-Image Generation Text-to-Image Generation	FRA-OOD Train / FRA-ID	1.4k 6k / 1.5k
Generated	IU	TextVQA ChartVQA InfographicsVQA	Text Reasoning Chart Reasoning Graph Reasoning	Train / FRA-ID Train / FRA-ID Train / FRA-ID	2.1k / 0.4 k 2.1k / 0.4 k 2.1k / 0.4 k
	IG	MagicBrush COCO	Image Editing Text-to-Image Generation	Train / FRA-ID Train / FRA-ID	3.5 k / 0.5 k 4.8k / 1k
	ITIG	VIST wikiHow InterleavedBench	Visual Story Completion Multimodal Script Generation Storytelling Generation, Activitynet Continuation	Train / FRA-ID Train / FRA-ID FRA-OOD	5.5k / 0.7k 6.4k / 0.6k 90

Figure 2: Pairwise data statistics of the FRAbench. "Public" indicates data derived from existing public datasets, "Generated" denotes data synthesized through our generation pipeline. Detailed counts of evaluation labels for each sub-task are in Appendix B.4.

Based on aspect trees, the fine-grained evaluation process can be streamlined, as illustrated in Figure 1.

3.2 Fine-grained Evaluation Dataset Construction

Based on the aspect tree, we can construct fine-grained evaluation datasets for training and testing UFEval. Specifically, we construct FRABench—a large-scale, multi-modal, and aspect-level evaluation dataset. Following prior studies (Kim et al., 2024a; Ye et al., 2024) that demonstrate pointwise scoring is more susceptible to contextual bias and available for reward model training, we adopt a pairwise comparison evaluation method in FRABench.

3.2.1 FRABench Construction.

We first collect questions from the datasets corresponding to the 28 sub-tasks we gathered, and generate pairwise samples. Specifically, we first obtain 29.3K response pairs from public datasets and generate the remaining 30.1K pairs using different MLLMs (Details regarding the MLLMs used can be found in Appendix B.1). After obtaining the queries and response pairs, we assign aspects from our aspects tree. As a result, each pairwise sample is annotated with an average of 8 UAs and 3 TAs (Detailed information on aspect assignment is provided in Appendix B.4). We then generate evaluation labels via two approaches: (1) Human annotations: directly using three-aspect human-annotated scores from ImageRewardDB (Xu et al., 2023a), supplemented with feedback generated by GPT-4o; (2) GPT-4o annotations: due to the lack of human annotations for most aspects, we use GPT-4o to generate evaluation labels for all other pairwise samples. During the GPT-4o annotation process, we observed that GPT frequently incorporates response correctness into its assessment of UAs. To address this, we provide only the response, excluding the query, when evaluating UAs. All templates are provided in Appendix G. Moreover, to mitigate position bias (Ye et al., 2024), we balance the number of samples where response 1 is preferred over response 2 and vice versa by reversing the response positions in half of the surplus samples from the majority class and re-annotating them accordingly (Appendix B.2 presents detailed analysis about position bias of FRABench). Finally, we generate 325k fine-grained evaluation labels.

3.2.2 Dataset Partitioning and Evaluator Training.

Due to the lack of a large-scale aspect-level benchmark to verify UFEval's aspect generalizable capability and support our arguments, we can only divide FRABench into training, In-Domain test set (FRA-ID), and Out-of-Domain test set (FRA-OOD). The training and FRA-ID consist of 18 randomly selected sub-tasks from four tasks, covering 22 UAs and 35 TAs, while FRA-OOD contains 10 unseen sub-tasks with 28 seen UAs and 27 unseen TAs. Finally, the training set contains 255.4K samples, FRA-ID contains 45.2K samples, and FRA-OOD contains 24.4K samples. The statistics of FRABench are presented in Figure 2, with more detailed analysis provided in Appendix B.

An evaluator's main job is to judge model responses like humans would, so their effectiveness depends on how closely their judgments match human opinions. To measure this, we create human-annotated evaluation datasets. Specifically, we extract partial samples from FRA-ID and FRA-

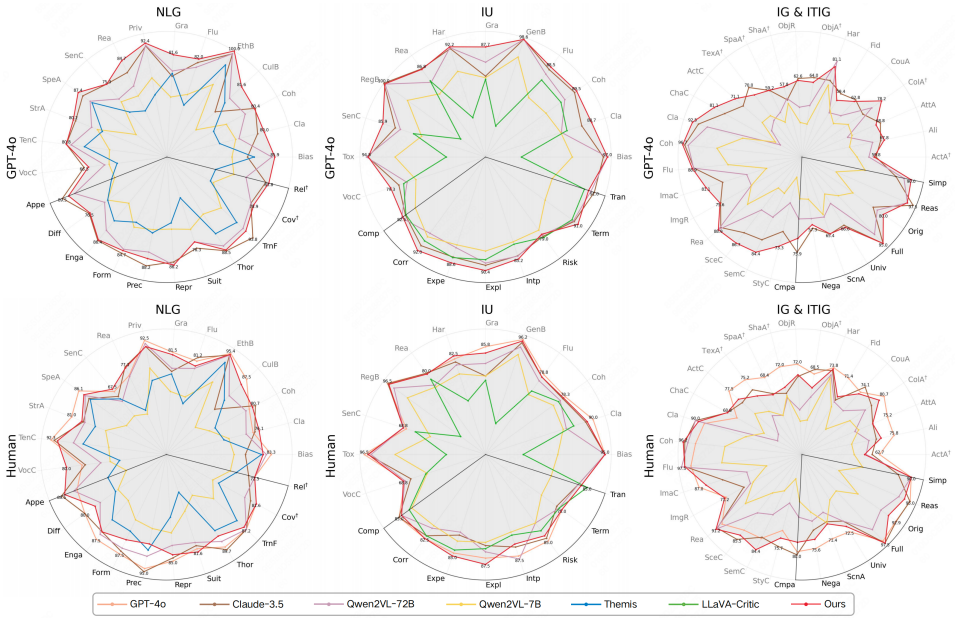


Figure 3: Comparison of baselines on the FRA-OOD and FRA-OOD-H. Black areas show unseen TAs (for aspect generalization), while the rest show seen UAs for unseen tasks (for task generalization).

Table 2: The results correspond to task and aspect generalization evaluation. Average accuracy serves as the evaluation metric. Bold and underline indicate the first and second best results, respectively.

Method	Task Generalization Evaluation								Aspect Generalization Evaluation							
	FRA-OOD (GPT4o)				FRA-OOD-H (Human)				FRA-OOD (GPT4o)				FRA-OOD-H (Human)			
	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG
GPT-4o	-	-	-	-	**84.0**	**82.1**	**72.3**	**93.1**	-	-	-	-	**83.2**	**82.1**	**74.2**	**93.1**
Claude-3.5	74.6	85.8	**72.5**	75.6	83.0	76.5	63.1	91.0	**84.1**	84.3	**65.6**	85.0	82.6	76.5	65.1	91.0
Qwen2VL-72B	70.2	82.4	65.8	60.0	78.3	75.3	48.6	83.7	76.3	81.5	**65.6**	85.0	77.3	75.3	53.8	83.7
Qwen2VL-7B	50.4	65.9	61.4	43.4	50.9	65.9	40.9	44.3	54.5	69.1	37.7	69.1	49.1	65.9	46.0	44.3
Themis	56.7	-	-	-	58.9	-	-	-	55.0	-	-	-	58.8	-	-	-
LLaVA-Critic	-	52.2	-	-	-	76.2	-	-	-	80.8	-	-	-	76.2	-	-
Ours	**81.7**	**90.4**	69.0	**83.1**	79.0	80.9	62.1	90.6	83.0	**86.3**	62.9	**89.3**	78.3	80.9	66.1	90.6

OOD, and recruited three humans with master's degrees for annotation. Finally, we retain 6.9K in-domain evaluation samples (FRA-ID-H) and 6.0K out-of-domain evaluation samples (FRA-OOD-H), respectively. Details of the human annotation process and annotation consistency are provided in Appendix E. When training UFEval, we use SFT to fine-tune Qwen2-VL-7B-Instruct on the training set. The detailed training configurations and methods are presented in Appendix C.1.

4 Experiments

To assess UFEval's generalizability from inter-aspect correlations, we first conduct evaluations on previously unseen tasks and aspects, termed Out-of-Domain Evaluation. Next, we evaluate UFEval as MLLM-as-a-Judge using public benchmarks. We further explore the effectiveness of Multi-aspect and Multi-task Assessment Learning for evaluators. Lastly, we validate UFEval's application in generating preference data for DPO-based alignment. In the following sections, we present our experimental setup and results (In-Domain Evaluation and other experimental results are in Appendix D).

4.1 Experimental Setup

4.1.1 Benchmarks.

(1) Out-of-Domain Evaluation: we use FRA-OOD and FRA-OOD-H to validate. (2) Evaluation as MLLM-as-a-Judge: we select public benchmarks across three tasks: For NLG, we use MT-

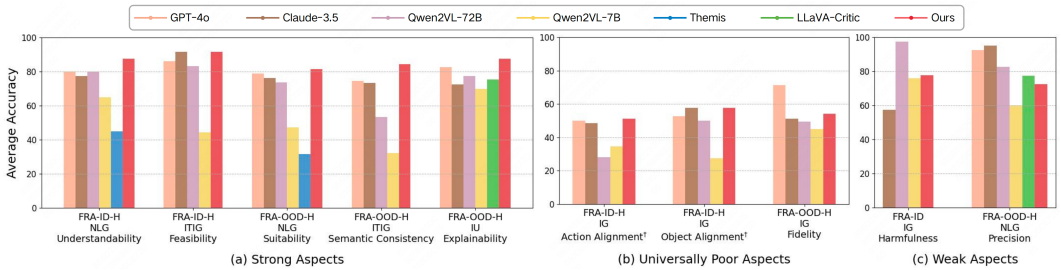


Figure 4: Comparison of aspect-level. In the subfigures, the x-axis shows the test set (FRA-ID, FRA-ID-H, FRA-OOD, or FRA-OOD-H) and task (NLG, IU, IG, or ITIG) of each evaluated aspect.

Table 3: Evaluation as MLLM-as-a-Judge for NLG, using three benchmarks. "tau" indicates that accuracy is calculated with ties, and "diff" excludes tied pairs when calculating accuracy.

Method	SummEval										MANS	MT-Bench	
	Coherence		Consistency		Fluency		Relevance		Ave.				
	tau(↑)	diff(↑)	tau(↑)	diff(↑)	tau(↑)	diff(↑)	tau(↑)	diff(↑)	tau(↑)	diff(↑)	diff(↑)	tau(↑)	diff(↑)
GPT-4o	58.0	64.2	79.1	85.1	64.3	72.8	60.1	67.1	65.3	72.3	68.5	70.9	83.5
Claude-3.5	63.5	70.6	81.6	**87.9**	73.7	83.1	59.6	66.6	**69.6**	**77.0**	68.4	**76.3**	**90.7**
Qwen2VL-72B	**66.8**	**73.2**	66.8	66.4	71.8	80.4	**62.2**	**69.3**	66.9	72.3	19.3	75.9	88.7
Qwen2VL-7B	48.8	52.5	35.5	30.6	44.8	49.8	41.7	44.5	42.7	44.3	60.2	44.5	50.1
Qwen3VL-8B	56.5	61.5	60.8	57.4	68.0	75.5	55.4	60.8	60.1	63.8	45.5	70.1	80.9
Themis	60.7	62.1	**81.8**	86.0	73.3	77.7	54.4	54.6	67.5	70.1	44.2	43.6	37.7
Auto-J	-	-	-	-	-	-	-	-	60.4	67.0	68.2	73.0	85.5
Prometheus 2	55.2	62.1	65.5	74.7	61.6	69.8	55.0	61.8	59.3	67.1	69.0	55.1	72.0
Ours	64.6	71.8	75.2	83.5	**74.7**	**84.5**	61.3	67.6	69.0	76.9	**69.3**	74.9	88.3

Bench (Zheng et al., 2023), SummEval (Fabbri et al., 2021), and MANS (Guan et al., 2021). Since SummEval and MANS only provide scores, we generate pairwise samples for evaluation through sample pairing. For IU, we use WildVision (Lu et al., 2024), MLLM-as-a-Judge (Chen et al., 2024), and VLRewardBench (Li et al., 2024e). For IG, GenAI-Bench (Li et al., 2024a), Winoground (Thrush et al., 2022), and Pick-a-Pic (Kirstain et al., 2023) are selected. To ensure fair comparison, we carefully check that the training set contains no overlapping samples with the benchmarks. (3) Multi-Aspect and Multi-Task Assessment Learning: We use FRA-ID to investigate the multi-aspect learning and benchmarks above to verify multi-task synergy. (4) Preference Alignment Comparison: We leverage UFEval for image generation and understanding model alignment. For image generation, we generate images using captions from the HPDv2 (Wu et al., 2023). For image understanding, we use MMHal (Sun et al., 2023), LLaVABen.Wild (Li et al., 2024b), and LLaVABen (Liu et al., 2023a). Detailed descriptions and usage of each benchmark are in Appendix C.2.

4.1.2 BASELINES.

We experiment with representative models from two categories: (1) Prompting Models: GPT-4o, Claude-3.5, Qwen3-VL-8B (Qwen3VL-8B), Qwen2-VL-72B-Instruct (Qwen2VL-72B), and Qwen2-VL-7B-Instruct (Qwen2VL-7B), implemented using the same prompts as UFEval. (2) Fine-tuned Evaluators: Themis-8B (Themis) (Hu et al., 2024b) , Auto-J (Li et al., 2023b), Prometheus 2 (Kim et al., 2024a), LLaVA-Critic-7B (LLaVA-Critic) (Xiong et al., 2024), ImageReward (Xu et al., 2023a), VisionReward (Xu et al., 2024), Q-Eval (Zhang et al., 2025) and CIGEval (Wang et al., 2025a). Notably, ImageReward and VisionReward use different evaluation paradigms from ours. Auto-J cannot assess specific aspects, and Prometheus 2 requires reference answers. Therefore, we use these methods only as baselines in the MLLM-as-a-Judge evaluation.

4.2 OUT-OF-DOMAIN EVALUATION

Tasks Generalization Evaluation. For task generalization, we employ the seen UAs and unseen tasks from UFEval during testing to control for test variables. The results are shown in the regions outside the black sectors in Figure 3 [2]. Detailed scores for each aspect are reported in Appendix D.4.

[2]We use aspect abbreviations here; the full names can be found in Appendix A.

Table 4: Evaluation as MLLM-as-a-Judge for IU. We evaluate baselines across three benchmarks.

Method	WildVision			MLLM-as-a-Judge		VLRewardBench			
	tau (↑)	diff (↑)	τ (↑)	tau (↑)	diff (↑)	General diff (↑)	Hallucination diff (↑)	Reasoning diff (↑)	Ave. diff (↑)
GPT-4o	**55.3**	**70.1**	**73.3**	58.1	67.0	50.2	81.4	**74.8**	**68.8**
Claude-3.5	53.3	67.3	61.2	**58.4**	**68.3**	38.5	**82.6**	66.1	62.4
Qwen2VL-72B	50.3	59.6	65.5	54.6	58.5	**50.8**	75.4	70.7	65.6
Qwen3VL-8B	48.1	57.7	54.5	53.0	57.5	45.5	66.0	69.0	60.1
LLaVA-Critic	53.0	66.0	59.6	55.6	65.5	42.0	41.2	60.0	47.7
Qwen2VL-7B	39.2	40.6	23.1	41.3	44.6	45.1	62.8	62.5	56.8
w/ IU.	47.3	60.1	58.3	50.1	59.0	45.6	57.6	68.5	57.2
w/ IU+ITIG.	51.2	65.2	64.1	54.1	64.8	46.5	56.9	70.7	58.0
w/ IU+IG.	52.7	67.9	66.0	56.1	66.8	46.4	57.0	71.0	58.1
Ours	53.9	68.6	66.5	57.2	67.0	46.4	57.7	71.1	58.4

Table 5: Evaluation as MLLM-as-a-Judge for IG. We evaluate baselines across three benchmarks.

Method	GenAI-Bench		Winoground				Pick-a-Pic	
	tau(↑)	diff(↑)	Relation diff (↑)	Object diff (↑)	Both diff (↑)	Ave. diff (↑)	tau(↑)	diff(↑)
GPT-4o	**55.6**	69.5	62.6	**73.0**	73.0	69.5	54.4	59.2
Claude-3.5	**55.6**	**71.0**	**71.2**	**73.0**	69.2	**71.1**	49.1	53.7
Qwen2VL-72B	49.1	52.6	46.7	60.9	57.6	55.0	38.6	38.3
Qwen3VL-8B	53.0	59.0	54.6	63.6	69.0	62.4	49.2	55.1
VisionReward	51.0	66.4	60.2	64.1	74.9	66.4	48.9	58.0
ImageReward	48.6	64.9	54.0	58.2	69.2	60.4	48.8	55.7
CIGEval	40.1	29.8	42.5	30.4	30.7	34.5	34.5	31.5
Q-Eval	48.0	65.5	62.4	54.5	61.5	59.4	**55.0**	**62.2**
Qwen2VL-7B	35.8	38.0	33.4	42.5	46.1	40.6	38.1	40.2
w/ IG.	46.6	59.4	53.1	52.0	71.2	58.7	45.8	53.1
w/ IG+ITIG.	50.1	62.6	56.0	58.2	74.2	62.8	47.1	55.9
w/ IG+IU.	52.5	64.5	57.3	58.0	78.5	64.6	49.5	56.9
Ours	53.6	65.5	57.5	59.1	**80.7**	65.7	50.0	57.3

UFEval demonstrates strong alignment with both GPT-4o and human annotators on all tasks. In the "Tasks Generalization Evaluation" column of Table 2, UFEval achieves overall average accuracies of 85% and 83% on FRA-OOD and FRA-OOD-H, respectively, and outperforms both Themis and LLaVA-Critic on all tasks. Supporting the effectiveness of aspect-level evaluation for cross-task generalization. We also provide several good and bad cases in Appendix H.

Aspects Generalization Evaluation. We use the unseen aspects of unseen tasks in FRA-OOD and FRA-OOD-H to evaluate UFEval's generalization to aspects. Specifically, this evaluation encompasses two distinct dimensions: (1) Contextual generalization, which validates whether similar semantics and standards can be transferred across diverse tasks (evaluated using 12 aspects); and (2) Novel aspect generalization, which assesses the capability to handle concepts semantically distinct from the training set (evaluated using 15 aspects). A detailed breakdown of these aspects is provided in Table 11. Moreover, to support the validity of aspect generalization evaluation, we provide ROUGE_L (Wang et al., 2022) tests on unseen aspects in Appendix B.3.

The results are shown in the black sectors of Figure 3. In terms of overall coverage, UFEval outperforms both Themis and LLaVA-Critic, the state-of-the-art evaluators in their respective domains. In the right column of Table 2, UFEval continues to demonstrate strong performance on the NLG, IU, IG and ITIG, with high consistency with both GPT-4o and human annotators, achieving overall accuracies of 86.2% and 83.2%, respectively. These results show that even without exposure to unseen aspects during training, UFEval remains effective in evaluation. This is largely due to the naturally related nature of aspects, which enables the transfer of similar meanings and standards.

Additionally, we sample three types of aspects for aspect-level analysis (Figure 4) from four test sets based on UFEval's performance: (1) Weak aspects, where the best model outperforms UFEval by over 20%; (2) Universally poor aspects, where all evaluators score very low; and (3) Strong aspects, where UFEval excels. Only two weak aspects occur in all test sets, and both are included; representative subsets are selected for the other categories.

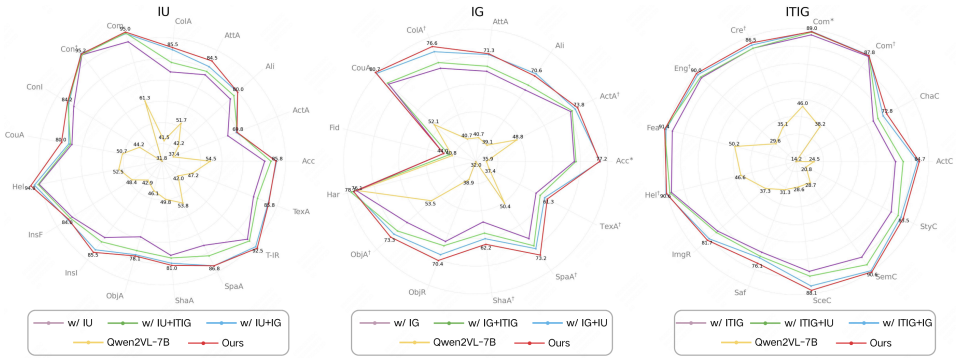


Figure 5: We evaluate multi-aspect learning by training models on different data combinations and testing them on three visual tasks (IU, IG, ITIG) in FRA-ID.

Regarding the weak aspects, as shown in Figure 4c, UFEval performs poorly on the harmfulness of IG in FRA-ID, mainly because it tends to classify shadowy or gloomy images as harmful due to heightened sensitivity to elements that may evoke psychological discomfort. In Figure 4b, all evaluators perform poorly on fidelity, which requires recognizing fine-grained object features (e.g., well-formed human facial features) that current models, including GPT-4o, often fail to detect (see Appendix H for more representative samples).

4.3 Evaluation as MLLM-as-a-Judge

For NLG Evaluations. The experimental results, shown in Table 3, demonstrate the effectiveness of UFEval in NLG tasks. Specifically, in SummEval and MT-Bench, UFEval outperforms most baselines, with Claude being the sole exception. In the MANA benchmark, UFEval achieves the best performance, with an accuracy of 69.3%. These results highlight UFEval's strong capability for text understanding evaluation compared to domain-specific models in this field.

For IU Evaluations. The results are presented in Table 4. Additionally, we assess model-level consistency on WildVision using Elo ratings and Kendall's Tau (τ) to compare model rankings (Xiong et al., 2024). The findings demonstrate that UFEval consistently outperforms LLaVA-Critic across all benchmarks. LLaVA-Critic focuses on respective datasets training, whereas UFEval uses joint learning across multiple tasks, which leads to relatively superior performance. For more comprehensive MLLM-as-a-Judge results, please refer to Table 49.

For IG Evaluations. The experimental results are shown in Table 5. Our evaluator outperforms ImageReward in all benchmarks, achieving accuracies of 65.5%, 65.7%, and 57.3%, respectively. Even compared to the state-of-the-art image evaluator VisionReward, the performance gap is minimal—only 0.9% lower on GenAI-Bench and 0.3% lower on Winogrounded. These results demonstrate the promising capability of UFEval in the image generation evaluation task.

4.4 Multi-aspect and Multi-task Assessment Learning

The construction of UFEval is based on our intuitive argument: jointly learning to assess multiple visual aspects and tasks may foster a synergistic effect. Therefore, we explore the effectiveness of multi-aspect and multi-task learning on the evaluators, respectively. Specifically, we experiment with various training data configurations to train the model, analyzing the influence of multi-aspect and multi-task learning. For instance, for IU tasks, we design three training configurations to investigate the impact of multi-aspect and multi-task training: (1) learning solely on IU aspect-level assessment (w /IU), (2) jointly learning IU and ITIG aspect-level assessment (w /IU+ITIG), and (3) jointly learning IU and IG aspect-level assessment (w /IU+IG).

Multi-task Assessment Learning. The results are presented in Table 4 and 5. Our results demonstrate that multi-task learning outperforms single-task training in enhancing evaluator performance. For instance, as shown in Table 4, when evaluating IU tasks, models trained jointly on both IU and IG tasks achieve superior overall accuracy compared to those trained exclusively on IU. These

Table 6: Image understanding DPO comparison. We compare our UFEval with LLaVA-Critic for DPO based on LLaVA-Next-7B.

Method	MMHal	LLaVABen.Wild	LLaVABen
LLaVA-Next-7B	2.05	52.9	28.3
w/ LLaVA-Critic	2.24	59.0	30.3
w/ UFEval	**2.41**	**61.4**	**32.3**

Table 7: Image generation DPO comparison. We evaluate DPO performance using UFEval-generated data versus the Pick-a-Pic dataset.

Method	HPSv2	ImageReward	VisionReward
SDXL	28.1	0.80	3.00
w/ Pick-a-Pic	28.7	0.84	3.20
w/ UFEval	**29.9**	**0.90**	**3.27**

findings underscore the significant advantages of exploiting shared representations and complementary knowledge across diverse visual tasks, ultimately yielding a more high-performing evaluator.

Multi-aspect Assessment Learning. The results presented in Figure 5 show that learning with multi-aspect from different tasks improves the UFEval's performance across most aspects. For example, in the IG task, incorporating relevant aspects from the IU, such as alignment in caption generation and recognition accuracy, enhances the evaluation of aspects like object alignment in IG. Similar improvements are observed in IU and ITIG. These results highlight the benefits of leveraging multi-aspect data to enrich the evaluator's understanding and shared aspect knowledge.

4.5 Preference Alignment Comparison

To validate the effectiveness of UFEval to generate preference data across both IU and IG tasks, we employ it to construct training data for DPO-based model alignment. The application principle of DPO for model alignment in both IU and IG tasks is comprehensively described in Appendix F.

For IU Tasks. Building upon UFEval, we use DPO to improve the image understanding capabilities of LLaVA-Next-7B (Li et al., 2024d). For fair comparison, both UFEval and the LLaVA-Critic ues identical image-question pairs sourced from RLHF-V (Yu et al., 2024) and LLaVA-RLHF (Sun et al., 2023) to construct preference data. Ultimately, we construct 15k preference samples. We then train LLaVA-Next-7B on 8 A100 GPUs using a batch size of 2, gradient accumulation steps of 2, a learning rate of 5×10^{-7}, and set β_u to 0.1. The results are shown in Table 6, UFEval consistently surpasses LLaVA-Critic across all evaluated benchmarks. Notably, it achieves a 2.4% improvement on LLaVABench.Wild, underscoring its enhanced effectiveness in visual understanding tasks.

For IG Tasks. Based on UFEval, we apply DPO to SDXL-Turbo (Podell et al., 2023), a conditional diffusion model, to better align its outputs with human preferences without explicit reward modeling. We extract prompt and two corresponding images from Pick-a-Pic and use UFEval to construct training preference data. In total, we construct 14k preference samples for DPO image generation. We then train SDXL-Turbo on 8 A100 GPUs using $\beta_g = 5000$ with a batch size of 32 for three epochs. HPSv2 (Wu et al., 2023), ImageReward, and VisionReward are used for quality assessment. The results are shown in Table 7, training on the constructed data using UFEval achieves better performance compared to directly training on the original dataset. The qualitative comparison results are shown in Appendix I. This demonstrates the effectiveness of our approach in refining preference data for improved model alignment in image generation tasks.

5 Conclusion, Limitation and Future Work

This paper proposes UFEval, the first unified fine-grained evaluator with task and aspect generalization. Specifically, we start by building a comprehensive aspect tree that leads to the creation of FRABench, a large-scale, multi-modal, and aspect-level evaluation dataset. We then fine-tune an MLLM on FRABench to develop UFEval. Our experimental results demonstrate that joint learning across diverse visual tasks and aspects yields significant mutual benefits and generalization capabilities. We also leverage UFEval to automatically construct high-quality preference pair datasets for DPO training to align models' outputs. These results validate the effectiveness of the FRABench as a valuable resource for training unified evaluators. However, compared to the other three tasks, UFEval has relatively limited performance in IG tasks. This may primarily be due to existing LMMs' insufficient active visual semantic understanding. In future work, we plan to incorporate video understanding and generation tasks into our evaluation system and add their corresponding aspects to the aspect tree.

ACKNOWLEDGMENTS

This project was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62576102.

REFERENCES

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, et al. Agentharm: A benchmark for measuring harmfulness of llm agents. arXiv preprint arXiv:2410.09024, 2024.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. Advances in Neural Information Processing Systems, 36:78142–78167, 2023.

Anna O Basile, Alexandre Yahi, and Nicholas P Tatonetti. Artificial intelligence for drug toxicity and safety. Trends in pharmacological sciences, 40(9):624–635, 2019.

João Bordalo, Vasco Ramos, Rodrigo Valério, Diogo Glória-Silva, Yonatan Bitton, Michal Yarom, Idan Szpektor, and Joao Magalhaes. Generating coherent sequences of visual illustrations for real-world manual tasks. arXiv preprint arXiv:2405.10122, 2024.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. ACM transactions on intelligent systems and technology, 15(3):1–45, 2024.

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

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. arXiv preprint arXiv:2310.02071, 2023.

Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In Proceedings of the 27th ACM International Conference on Multimedia, pp. 2236–2244, 2019.

Daniel DeAlcala, Ignacio Serna, Aythami Morales, Julian Fierrez, and Javier Ortega-Garcia. Measuring bias in ai models: an statistical approach introducing n-sigma. In 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), pp. 1167–1172. IEEE, 2023.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 862–872, 2021.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409, 2021.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. arXiv preprint arXiv:2302.04166, 2023.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. Computational Linguistics, 50(3):1097–1179, 2024.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu (eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 3356–3369, Online, November 2020. doi: 10.18653/v1/2020.findings-emnlp.301.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems, 36: 52132–52152, 2023.

Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Jian Guan, Zhexin Zhang, Zhuoer Feng, Zitao Liu, Wenbiao Ding, Xiaoxi Mao, Changjie Fan, and Minlie Huang. Openmeva: A benchmark for evaluating open-ended story generation metrics. arXiv preprint arXiv:2105.08920, 2021.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. Bias in large language models: Origin, evaluation, and mitigation. arXiv preprint arXiv:2411.10915, 2024.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. Evaluating large language models: A comprehensive survey. arXiv preprint arXiv:2310.19736, 2023.

Jeffrey Heinz and William Idsardi. Sentence and word complexity. Science, 333(6040):295–297, 2011.

Or Honovich, Thomas Scialom, Omer Levy, and Timo Schick. Unnatural instructions: Tuning language models with (almost) no human labor. arXiv preprint arXiv:2212.09689, 2022.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR), 51(6):1–36, 2019.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. Are llm-based evaluators confusing nlg quality criteria? arXiv preprint arXiv:2402.12055, 2024a.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. Themis: A reference-free nlg evaluation language model with flexibility and interpretability. arXiv preprint arXiv:2406.18365, 2024b.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems, 36:78723–78747, 2023.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1–55, 2025.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. Tigerscore: Towards building explainable metric for all text generation tasks. arXiv preprint arXiv:2310.00752, 2023.

Pei Ke, Bosi Wen, Zhuoer Feng, Xiao Liu, Xuanyu Lei, Jiale Cheng, Shengyuan Wang, Aohan Zeng, Yuxiao Dong, Hongning Wang, et al. Critiquellm: Towards an informative critique generation model for evaluation of large language model generation. arXiv preprint arXiv:2311.18702, 2023.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. Prometheus: Inducing fine-grained evaluation capability in language models. In The Twelfth International Conference on Learning Representations, 2023.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. arXiv preprint arXiv:2405.01535, 2024a.

Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–21, 2024b.

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.

Mahnaz Koupaee and William Yang Wang. Wikihow: A large scale text summarization dataset. arXiv preprint arXiv:1810.09305, 2018.

Isack Lee and Haebin Seong. Do llms have political correctness? analyzing ethical biases and jailbreak vulnerabilities in ai systems. arXiv preprint arXiv:2410.13334, 2024.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, et al. Genai-bench: Evaluating and improving compositional text-to-visual generation. arXiv preprint arXiv:2406.13743, 2024a.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024b. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems, 36: 28541–28564, 2023a.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, Jianfeng Gao, et al. Multimodal foundation models: From specialists to general-purpose assistants. Foundations and Trends® in Computer Graphics and Vision, 16(1-2):1–214, 2024c.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024d.

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

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, et al. M^3 it: A large-scale dataset towards multi-modal multilingual instruction tuning. arXiv preprint arXiv:2306.04387, 2023c.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, et al. Vlrewardbench: A challenging benchmark for vision-language generative reward models. arXiv preprint arXiv:2411.17451, 2024e.

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. arXiv preprint arXiv:2410.09421, 2024f.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. Exploring the reliability of large language models as customized evaluators for diverse nlp tasks. arXiv preprint arXiv:2310.19740, 2023d.

Zhichao Liao, Xiaokun Liu, Wenyu Qin, Qingyu Li, Qiulin Wang, Pengfei Wan, Di Zhang, Long Zeng, and Pingfa Feng. Humanaesexpert: Advancing a multi-modality foundation model for human image aesthetic assessment. arXiv preprint arXiv:2503.23907, 2025.

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

Minqian Liu, Ying Shen, Zhiyang Xu, Yixin Cao, Eunah Cho, Vaibhav Kumar, Reza Ghanadan, and Lifu Huang. X-eval: Generalizable multi-aspect text evaluation via augmented instruction tuning with auxiliary evaluation aspects. arXiv preprint arXiv:2311.08788, 2023b.

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, 2024a.

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. Hd-eval: Aligning large language model evaluators through hierarchical criteria decomposition. arXiv preprint arXiv:2402.15754, 2024b.

Yujie Lu, Dongfu Jiang, Wenhu Chen, William Yang Wang, Yejin Choi, and Bill Yuchen Lin. Wildvision: Evaluating vision-language models in the wild with human preferences. arXiv preprint arXiv:2406.11069, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244, 2022.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 1697–1706, 2022.

Birger Moell and Johan Boye. Language complexity measurement as a noisy zero-shot proxy for evaluating llm performance. arXiv preprint arXiv:2502.11578, 2025.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734, 2021.

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.

Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. arXiv preprint arXiv:2501.18099, 2025.

Emily Sheng, Josh Arnold, Zhou Yu, Kai-Wei Chang, and Nanyun Peng. Revealing persona biases in dialogue systems. arXiv preprint arXiv:2104.08728, 2021.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8317–8326, 2019.

Marko Smilevski, Ilija Lalkovski, and Gjorgji Madjarov. Stories for images-in-sequence by using visual and narrative components. In ICT Innovations 2018. Engineering and Life Sciences: 10th International Conference, ICT Innovations 2018, Ohrid, Macedonia, September 17–19, 2018, Proceedings 10, pp. 148–159. Springer, 2018.

Saku Sugawara, Yusuke Kido, Hikaru Yokono, and Akiko Aizawa. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 806–817, 2017.

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.

Maurice C Taylor and MW Wahlstrom. Readability as applied to an abe assessment instrument. In PUB DATE Jun 84 NOTE 404p.; For other proceedings, see ED 299 461, CE 062 030-032, and CE 062 035-037. PUB TYPE Collected Works Conference Proceedings (021) LANGUAGE English; French, pp. 15. ERIC, 1999.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5238–5248, 2022.

Alakananda Vempala and Daniel Preoţiuc-Pietro. Categorizing and inferring the relationship between the text and image of twitter posts. In Proceedings of the 57th annual meeting of the Association for Computational Linguistics, pp. 2830–2840, 2019.

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, pp. 8228–8238, 2024.

Jifang Wang, Yangxue Yangxue, Longyue Wang, Zhenran Xu, Yiyu Wang, Yaowei Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, and Min Zhang. A unified agentic framework for evaluating conditional image generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12626–12646. Association for Computational Linguistics, July 2025a. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.620. URL https://aclanthology.org/2025.acl-long.620/.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. Shepherd: A critic for language model generation. arXiv preprint arXiv:2308.04592, 2023a.

Weichuan Wang, Zhaoyi Li, Defu Lian, Chen Ma, Linqi Song, and Ying Wei. Mitigating the language mismatch and repetition issues in llm-based machine translation via model editing. arXiv preprint arXiv:2410.07054, 2024.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. arXiv preprint arXiv:2503.05236, 2025b.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. arXiv preprint arXiv:2306.05087, 2023b.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. arXiv preprint arXiv:2212.10560, 2022.

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.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. Llava-critic: Learning to evaluate multimodal models. arXiv preprint arXiv:2410.02712, 2024.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. Advances in Neural Information Processing Systems, 36:15903–15935, 2023a.

Jiazheng Xu, Yu Huang, Jiale Cheng, Yuanming Yang, Jiajun Xu, Yuan Wang, Wenbo Duan, Shen Yang, Qunlin Jin, Shurun Li, et al. Visionreward: Fine-grained multi-dimensional human preference learning for image and video generation. arXiv preprint arXiv:2412.21059, 2024.

Wenda Xu, Danqing Wang, Liangming Pan, Zhenqiao Song, Markus Freitag, William Yang Wang, and Lei Li. Instructscore: Explainable text generation evaluation with finegrained feedback. arXiv preprint arXiv:2305.14282, 2023b.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, pp. 100211, 2024.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. arXiv preprint arXiv:2410.02736, 2024.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. arXiv preprint arXiv:2307.10928, 2023.

Jiahao Ying, Yixin Cao, Yushi Bai, Qianru Sun, Bo Wang, Wei Tang, Zhaojun Ding, Yizhe Yang, Xuanjing Huang, and Shuicheng Yan. Automating dataset updates towards reliable and timely evaluation of large language models. Advances in Neural Information Processing Systems, 37: 17106–17132, 2024.

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, pp. 13807–13816, 2024.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. arXiv preprint arXiv:2310.07641, 2023.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in Neural Information Processing Systems, 36:31428–31449, 2023a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415, 2023b.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. arXiv preprint arXiv:2306.17107, 2023c.

Zicheng Zhang, Tengchuan Kou, Shushi Wang, Chunyi Li, Wei Sun, Wei Wang, Xiaoyu Li, Zongyu Wang, Xuezhi Cao, Xiongkuo Min, et al. Q-eval-100k: Evaluating visual quality and alignment level for text-to-vision content. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 10621–10631, 2025.

Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. Svit: Scaling up visual instruction tuning. arXiv preprint arXiv:2307.04087, 2023.

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.

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

A Aspect Sources and Definitions

We provide the sources and definitions of aspects under two categories: UAs and TAs, as shown in Tables 13 to 23. For UAs, the "Target" column indicates the applicable output modality: "T" for Text, "I" for Image, and "ITI" for Interleaved Text-with-Image. For TAs, the "Sub-Task" column specifies the corresponding task type. Additionally, underscores are used to benchmark extended TAs for ITIG. Abbreviations for all aspects are provided in the table.

B Analysis of the FRABench

In this section, we provide detailed information and comprehensive analysis about the FRABench.

Table 8: The detailed statistical information of the sub-tasks, which is not shown in Figure 2.

Task	Sub-Task	Split	Size	Task	Sub-Task	Split	Size
NLG (Public)	Summarization	Train	218	IU (Public)	Detailed Image Captioning	Train	2.1k
		Test	54			Test	0.4k
	Creative Writing	Train	575		Robustness-oriented Instructions	Train	4k
		Test	120			Test	0.4k
	Rewriting	Train	194		Medical Image Understanding	Train	2.1k
		Test	26			Test	0.4k
	General Communication	Train	1080		Text-rich Understanding	Train	2.1k
		Test	263			Test	0.4k
	Functional Writing	Train	433		General Visual Conversation	Train	2.1k
		Test	147			Test	0.4k
	Question Generation	Test	132		Simple Image Captioning	Train	2.1k
						Test	0.4k
	Title Generation	Test	38		Embodied Decision-making	Test	0.4k
	Keywords Extraction	Test	80		Medical VQA	Test	0.4k
	Data Analysis	Test	170		Academic VQA	Test	0.4k
	Translation	Test	130	ITIG (Generated)	Storytelling Generation	Test	50
					Activity Continuation	Test	40

Table 9: Different MLLMs used for generating pairwise responses for sub-tasks under the "Generated" column in Figure 2 of the main text.

Task	Sub-Task	Model
IU	Text Reasoning Chart Reasoning Graph Reasoning	LLaVA-1.5-13B, Qwen2-VL-72B, InstructBLIP-Vicuna-13B, InternVL2-26B, Molmo-7B-D
IG	Text-to-Image Generation	Show-o-1.3B, Seed-X-17B, Flux-12B, Stable-Diffusion-3.5-Large, Ground Truth
	Image Editin	MagicBrush, SEED-X-17, InsPix2Pix, MGIE, Ground Truth
ITIG	Visual Story Completion Multimodal Script Generation Storytelling Generation Activitynet Continuation	miniGPT, GILL, MM-Interleaved, GPT4o+SDXL, Ground Truth

B.1 More Detailed Information about the FRABench

We provide the information about sub-tasks for NLG and IU in Table 8, as this information is not shown in Figure 2 of the main text. For the tasks in the "Generated" column in Figure 2 of the main text, due to the absence of pairwise data, we use four MLLMs with varying performance levels to generate pairwise responses for each query, as illustrated in Table 9.

B.2 Position Bias Analysis

We provide the counts of evaluation samples in the training set, FRA-ID, FRAUAs-OOD, and FRA-OOD, showing cases where Response 1 is better than, equal to, or worse than Response 2, as illustrated in Figures 6 to 9. As shown in Figure 6, the four tasks in the training set maintain a relatively consistent distribution. In Table 10, we also evaluate the position consistency of UFEval on the FRA-ID and FRA-OOD. We perform two inferences on the same sample, swapping the positions of the responses in the second inference, and calculate the average consistency, i.e., whether the results from the two inferences align. These experimental results demonstrate that UFEval achieves higher positional consistency compared to other baselines.

B.3 Aspect Diversity Analysis

Following prior work (Wang et al., 2022; Honovich et al., 2022), we analyze the ROUGE-L distribution between aspects in the training set and the FRA-OOD, respectively. Specifically, we sample

Table 10: We calculate the models' position consistency in the FRA-ID and FRA-OOD. Since our evaluator is trained on a relatively balanced dataset, it can mitigate the impact of position bias.

Method	FRA-ID				FRA-OOD			
	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG
GPT-4o	81.4	85.2	78.3	81.9	76.3	83.0	75.7	86.0
Claude-3.5	79.9	75.1	77.4	85.5	74.9	72.1	75.4	88.5
Qwen2VL-72B	76.0	69.9	77.8	75.2	76.2	78.1	70.0	86.6
Qwen2VL-7B	29.7	27.4	62.4	19.6	24.5	30.9	46.5	20.5
LLaVA-Critic	-	82.7	-	-	-	84.7	-	-
Ours	82.1	82.6	75.8	83.8	78.9	88.2	75.5	89.8

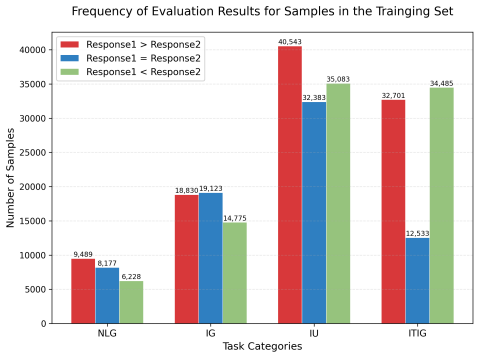


Figure 6: Statistics of pairwise data in the training set.

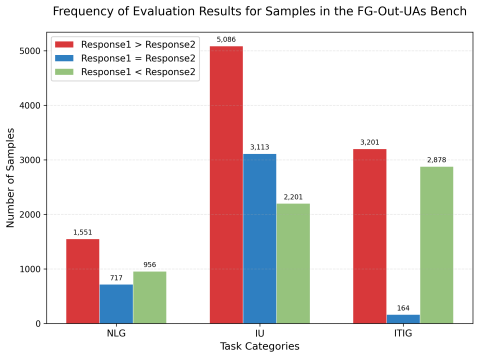


Figure 7: Statistics of pairwise data in the FRAUAs-OOD.

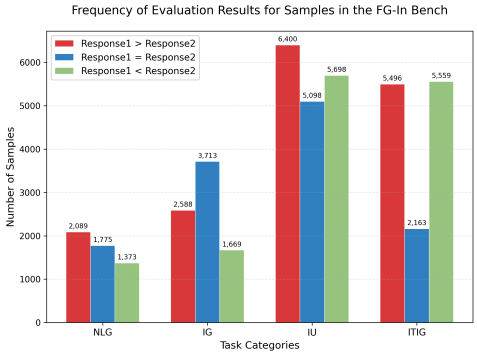


Figure 8: Statistics of pairwise data in the FRA-ID.

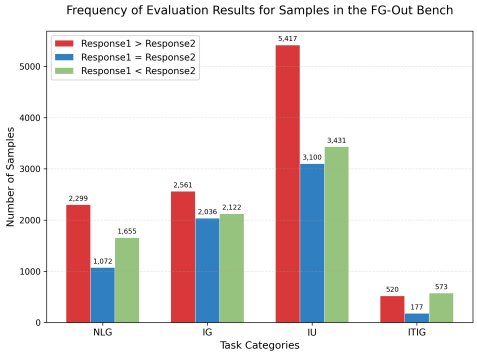


Figure 9: Statistics of pairwise data in the FRA-OOD.

pairs of aspects from the training set and compute their ROUGE-L scores. The overall distribution is shown in Figure 10, which indicates that the selected aspects are distinct from one another, confirming the inclusion of diverse and novel aspects in the training set. Additionally, we measure the ROUGE-L scores between aspects by sampling one aspect from the training set and another from the FRA-OOD. As shown in Figure 11, aspects in the training set and those in the FRA-OOD do not overlap, demonstrating the validity of our out-of-domain evaluations.

B.4 Aspect Selection Across Different Sub-tasks

We provide detailed information on the aspects selected for the training set, FRA-ID, FRAUAs-OOD, and FRA-OOD, as presented in Table 12. Additionally, we include statistics regarding the number of evaluation labels used for training and testing in each sub-task. Blue-marked aspects are human-

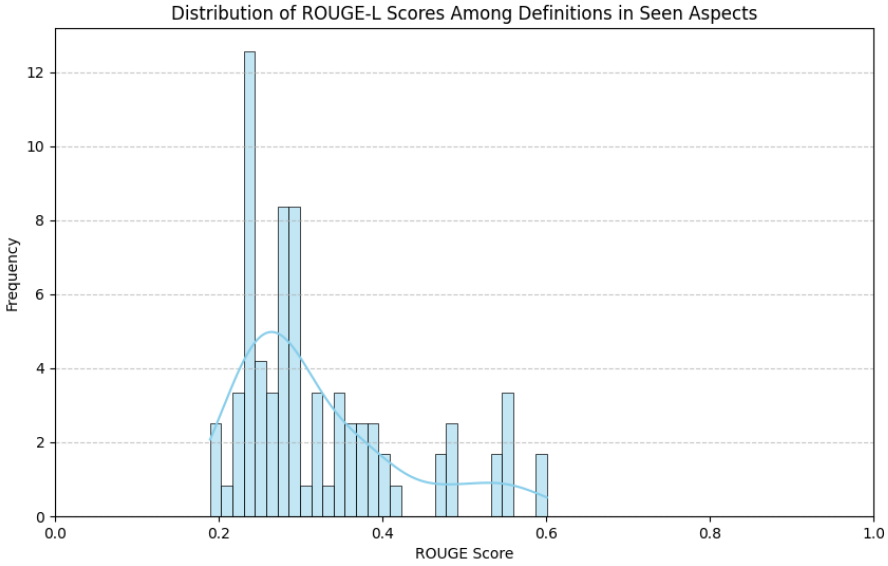


Figure 10: Rouge-L score distribution among two randomly sampled aspects from the training set. A left-skewed distribution with low values shows that the aspects are diverse.

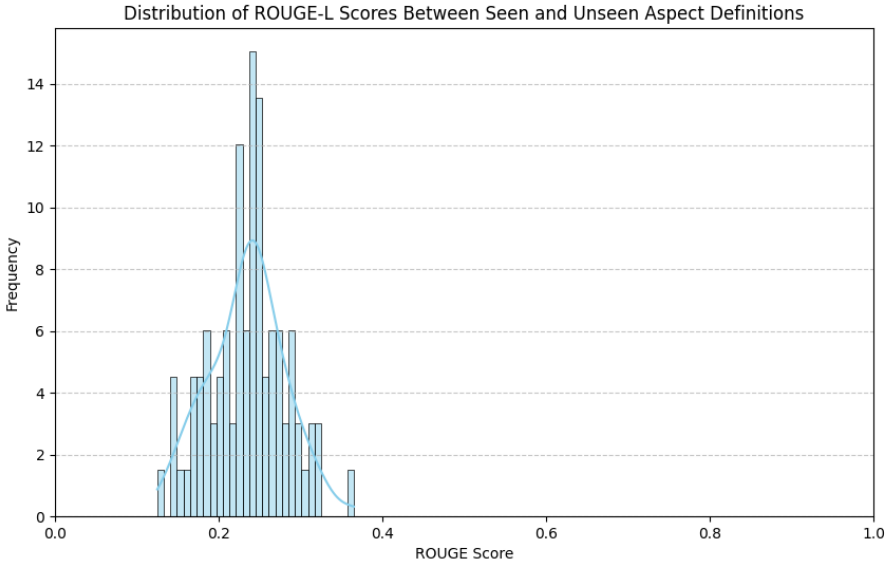


Figure 11: Rouge-L score distribution among a randomly sampled aspect from the training set and an aspect from the FRA-OOD. A left-skewed distribution with low values shows that they do not overlap with each other, hence meaning that the unseen aspect assumption is satisfied.

annotated, while the rest are annotated by GPT-4o. The human-annotated data in IG is sourced from the ImageRewardDB dataset (Xu et al., 2023a).

C EXPERIMENT DETAIL

C.1 FINE-TUNING DETAILS OF UFEVAL

In our experiments, we fine-tune Qwen2-VL-7B-Instruct on the training dataset using the Llama-Factory framework. We use 64 NVIDIA A100 GPUs, where each GPU processes one sample, and employ gradient accumulation with a factor of 2. The total batch size is 128. Additionally, we offload the optimizer parameters to the CPU to reduce GPU memory usage. The training process is carried out over 3 epochs. We set the learning rate to 1e-5, with 1% warmup steps. The learning rate follows a cosine decay schedule, gradually reducing as training progresses. UFEval is trained to predict pairwise rankings based on the criteria in the evaluation prompt, and provide detailed justifications for the assigned judgments. Standard cross-entropy loss is applied to both judgments and justifications.

C.2 DETAILED INFORMATION OF PUBLIC BENCHMARKS

We provide detailed descriptions and experimental information for the public benchmarks:

- SummEval (Fabbri et al., 2021): SummEval evaluates LLMs on text summarization across four aspects: Fluency, Consistency, Coherence, and Relevance. The dataset contains 100 instructions per aspect, each with responses from 16 different LLMs and corresponding human ratings (totaling 6.4k annotated samples). For testing, we randomly generate 9 comparison pairs from the 16 responses per instruction, creating 3.6k evaluation samples.
- MANS (Guan et al., 2021): MANS evaluates LLMs on the story generation task. Each instruction includes 5 different responses with human annotation, totaling 400 samples. For each instruction, we randomly select two responses to form pairwise data, resulting in 400 evaluation samples.
- MT-Bench (Zheng et al., 2023): MT-Bench is specifically designed to evaluate LLMs' capability as evaluators. The benchmark incorporates pairwise comparison data for 80 questions. Each pairwise data receives multiple human annotations, resulting in a total of 3k votes. To ensure high-quality evaluation samples, we select only those cases where all human annotators reach complete consensus, ultimately obtaining 684 evaluation samples.
- WildVision (Lu et al., 2024): WildVision comprises 11k human-annotated preference relations among LMM response pairs. Each relation includes a question-image pair and two responses generated by different models, with a human-annotated preference (including ties). For testing, we randomly select 2k response pairs from the dataset using the same testing protocol as LLaVA-Critic.
- MLLM-as-a-Judge (Chen et al., 2024): MLLM-as-a-Judge establishes a novel framework for assessing how closely model evaluations align with human judgments. The benchmark aggregates 17k multimodal evaluation instances (image-instruction-response triplets) spanning 14 diverse benchmarks and incorporating outputs from 6 different MLLMs. Through systematic pairwise comparisons conducted by human evaluators, the dataset provides 5,719 carefully annotated judgment cases for analysis.
- VLRewardBench (Li et al., 2024e): VLRewardBench is a comprehensive benchmark spanning general multimodal queries, visual hallucination detection, and complex reasoning tasks. Through an AI-assisted annotation pipeline combining sample selection with human verification, it curates 1,250 high-quality examples specifically designed to probe model limitations.
- GenAI-Bench (Li et al., 2024a): GenAI-Bench is a comprehensive benchmark with 1.6k compositional prompts to evaluate text-to-visual generation, surpassing the size and difficulty of existing benchmarks. Additionally, it provides over 15k human ratings for multiple aspects to support research on vision-language alignment metrics.
- Winoground (Thrush et al., 2022): Winoground is a benchmark for evaluating vision-language models' ability to perform visio-linguistic compositional reasoning—the capacity to understand how the meaning of an image changes when the same words are rearranged in a caption. It was hand-curated by expert annotators and labeled with fine-grained tags, containing 400 samples.
- Pick-a-Pic (Kirstain et al., 2023): Pickapic is a large dataset of text-to-image prompts and real users' preferences over generated images. Authors create a web app that enables text-to-image users to generate images and specify their preferences, which constructs 500 test samples.

Table 11: Aspect generalization performance evaluated across two dimensions: Contextual Generalization and Novel Aspect Generalization.

Sub-Task	Contextual Generalization	Novel Aspect Generalization
Question Generation	Engagingness	Difficulty
Title Generation	Coverage, Appeal	Suitability
Keywords Extraction	Relevance	Representation
Data Analysis	Precision, Format	-
Translation	-	Translation Fidelity, Thoroughness
Academic VQA	Correctnes	Expertise, Explainability
Medical VQA	-	Terminology, Transparency
Embodied Decision	-	Interpretability, Completion, Riskiness
Text-to-Image	Scene Alignment, Negatiion Universal, Comparison	- -
Storytelling Generation	Originality	Simplification
ActivityNet	-	Reasonableness, Fulfillment

- HPDv2 (Wu et al., 2023): HPDv2 is a large-scale (798k preference choices / 430k images), well-annotated dataset of human preference choices on images generated by text-to-image generative models. We utilized 400 prompts from the test set included in HPDv2 to generate corresponding images using SDXL-Turbo, and subsequently evaluated the generated images using three different quality assessment methods.
- MMHal (Sun et al., 2023): MMHal-Bench is an evaluation benchmark specifically designed for hallucination in Large Multimodal Models (LMMs). It contains 96 challenging questions based on images from OpenImages, and their corresponding ground-truth answers and image contents.
- LLaVABen (Liu et al., 2023a): LLaVA-Bench (in the wild) comprises 60 tasks designed to test visual instruction-following and question-answering capabilities in natural settings. Each task is scored by GPT-4 based on the correctness of the model's response relative to GPT-4-generated ground truth, with scores ranging from 0 to 1, aggregated across 60 samples.
- LLaVABen.Wild (Li et al., 2024b): LLaVA-Bench (Wilder) is an expanded benchmark for evaluating visual chat capabilities of MLLMs. It offers a compact 120-example set for quick evaluation. The dataset covers diverse real-world scenarios, including mathematical problem-solving, image comprehension, code generation, visual AI assistance, and image-based reasoning.

C.3 Aspects Used in the public benchmark

Since UFEval requires specific aspects for evaluation, we carefully assign one evaluation aspect to each benchmark. Specifically, we select suitable aspects from TAs and UAs based on the characteristics of each task. For benchmarks involving unseen tasks that may not align with existing aspects, we define new appropriate aspects based on their unique task properties. Tables 24 to 27 present the complete mapping of aspects used by UFEval across all benchmarks, with blue color indicating unseen aspects that were newly defined. To ensure fair comparison, GPT-4o, Claude-3.5, Qwen2VL-7B, and Qwen2VL-72B also follow the same aspect assignments when evaluated.

Table 12: Selecting Aspects and sample counts in the Training Set, FRA-ID, FRAUAs-OOD, and FRA-OOD. Blue-marked aspects are human-annotated, while the rest are annotated by GPT-4o.

DS	Sub-Task	Aspect Split	Universal Aspect	Specific-task Aspect	Size
Training Set & FRA-ID	Summarization	Train / Test	Clarity, Grammaticality, Coherence, Fluency	Coverage, Length Constraint, Layout, Conciseness	1.7k / 0.4k
	Rewriting	Train / Test	Grammaticality, Coherence, Fluency, Clarity	Accuracy, Pointing Out, Difference	1.3k / 0.1k
	Creative Writing	Train / Test	Gender Bias, Ethnic Bias, Regional Bias Harmfulness, Grammaticality, Readability	Creativity, Engagingness, Theme	5.1k / 1k
	Functional Writing	Train / Test	Clarity, Coherence, Fluency Gender Bias, Ethnic Bias	Relevance, Example Quality Contextual Information	3.4k / 1.1k
	General Communication	Train / Test	Gender Bias, Ethnic Bias, Regional Bias Clarity, Coherence, Fluency	Accuracy, Information Richness Understandability	11.8k / 2.6k
	General Visual Conv	Train / Test	Clarity, Harmfulness	Helpfulness, Instruction Following	8.4k / 1.6k
	Text-rich Und	Train / Test	Harmfulness, Vocabulary Complexity, Clarity	Accuracy, Instruction Following	10.5k / 2k
	Robustness-oriented Instructions	Train / Test	Spelling Accuracy, Structure Accuracy, Privacy Violations	Instruction Inconsistency, Context Inconsistency	19.6k / 2.5k
	Medical Image Und	Train / Test	Vocabulary Complexity, Bias, Sentence Complexity	Accuracy	8.3k / 2k
	Chart Reasoning	Train / Test	-	Accuracy, Instruction Following	4.2k / 0.8k
	Text Reasoning	Train / Test	-	Accuracy, Instruction Following	4.2k / 0.8k
	Graph Reasoning	Train / Test	-	Accuracy, Instruction Following	4.2k / 0.8k
	Simple Image Captioning	Train / Test	Readability, Bias	Conciseness, Object Alignment Text-Image Relationship, Alignment	12.5k / 2.4k
	Detailed Image Captioning	Train / Test	Coherence, Fluency, Clarity, Bias	Completeness, Attribute Alignment, Spatial Alignment Count Alignment, Action Alignment, Object Alignment Color Alignment, Texture Alignment, Shape Alignment	27.2k / 5.2k
	Text-to-Image Generation (COCO&DB)	Train / Test	Harmfulness, Fidelity	Count Alignment, Object Relationship Texture Alignment, Attribute Alignment, Object Alignment, Spatial Alignment, Alignment Action Alignment, Color Alignment, Shape Alignment	49.2k / 7.5k
	Image Editing	Train / Test	-	Accuracy	3.5k / 0.5k
	Visual Story Completion	Train / Test	Scene Consistency, Bias, Toxicity, Character Consistency Stylistic Consistency, Action Consistency, Fluency	Creativity, Engagingness, Completeness	39.8k / 7.7k
	Multimodal Script Generation	Train / Test	Action Consistency, Coherence, Image Repetition Readability, Semantic Consistency, Stylistic Consistency	Completeness, Feasibility Helpfulness, Safety	40.5k / 6.2k
FRAUAs-OOD	Summarization	Test	Sentence Complexity, Vocabulary Complexity, Readability Spelling Accuracy, Structure Accuracy, Text Repetition	-	1.3k
	Rewriting	Test	Tense Consistency, Spelling Accuracy Structure Accuracy, Text Repetition, Readability	-	0.9k
	Creative Writing	Test	Coherence, Fluency, Clarity, Text Repetition Sentence Complexity, Vocabulary Complexity, Cultural Bias	-	3.4k
	Functional Writing	Test	Readability, Spelling Accuracy, Structure Accuracy Sentence Complexity, Vocabulary Complexity	-	2.1k
	General Communication	Test	Readability, Harmfulness, Sentence Complexity Vocabulary Complexity, Grammaticality	-	5.4k
	General Visual Conv	Test	Fluency, Coherence, Grammaticality, Readability	-	1.6k
	Text-rich Und	Test	Fluency, Coherence, Grammaticality, Readability	-	1.6k
	Robustness-oriented Instructions	Test	Readability, Bias, Fluency, Coherence	-	1.6k
	Medical Image Und	Test	Fluency, Coherence, Readability, Grammaticality	-	1.6k
	Simple Image Captioning	Test	Fluency, Clarity, Ethnic Bias, Coherence, Sentence Complexity, Vocabulary Complexity	-	2.4k
	Detailed Image Captioning	Test	Readability, Grammaticality Sentence Complexity, Vocabulary Complexity	-	1.6k
	Visual Story Completion	Test	Readability, Coherence, Grammaticality Image Repetition, Semantic Consistency	-	3.8k
	Multimodal Script Generation	Test	Grammaticality, Fluency Scene Consistency, Clarity, Image Coherence	-	2.4k
FRA-OOD	Question Generation	Test	Clarity, Grammaticality, Fluency, Coherence Bias, Structure Accuracy, Spelling Accuracy	Engagingness, Difficulty	1.1k
	Title Generation	Test	Readability, Clarity, Bias, Ethical Bias Coherence, Fluency, Grammaticality, Cultural Bias	Suitability, Coverage, Appeal	0.4k
	Keywords Extraction	Test	Readability, Grammaticality Clarity, Fluency, Coherence	Relevance, Representation	0.5k
	Data Analysis	Test	Coherence, Privacy Violations, Readability, Fluency Clarity, Vocabulary Complexity, Sentence Complexity	Precision, Format	1.5k
	Translation	Test	Structure Accuracy, Spelling Accuracy, Tense Consistency Fluency, Grammaticality, Clarity, Coherence	Translation Fidelity, Thoroughness	1.1k
	Academic VQA	Test	Fluency, Readability, Vocabulary Complexity Sentence Complexity, Coherence, Grammaticality	Expertise, Correctness, Explainability	3.6k
	Medical VQA	Test	Sentence Complexity, Readability, Vocabulary Complexity Fluency, Coherence, Grammaticality, Clarity	Terminology, Transparency	3.6k
	Embodied Decision-makin	Test	Coherence, Clarity, Grammaticality, Gender Bias Harmfulness, Bias, Readability, Regional Bisa, Toxicity	Interpretability, Completion, Riskiness	4.8k
	Text-to-Image Generation (GenAI-Bench)	Test	Harmfulness, Fidelity	Count Alignment, Object Relationship Scene Alignment, Negation, Universal Object Alignment, Spatial Alignment, Alignment Action Alignment, Color Alignment, Shape Alignment Texture Alignment, Attribute Alignment, Comparison	6.7k
	Activitynet Continuation	Test	Readability, Harmfulness, Coherence, Semantic Consistency Stylistic Consistency, Action Consistency, Fluency, Clarity Scene Consistency, Image Repetition, Character Consistency	Reasonableness, Fulfillment	0.5k
	Storytelling Generation	Test	Readability, Harmfulness, Scene Consistency, Image Coherence Stylistic Consistency, Action Consistency, Semantic Consistency Image Repetition, Character Consistency, Fluency, Coherence	Originality, Simplification	0.6k

Table 13: The definition and source of Universal Aspects (UAs). (Part 1)

Aspect	Abbr	Source	Target	Definition
Bias	Bias	(Gallegos et al., 2024; Guo et al., 2024; Chang et al., 2024; Guo et al., 2024)	T	Bias evaluates the presence and extent of gender, regional, and cultural bias in generated content. It examines whether the content reinforces stereotypes, shows favoritism or discrimination toward specific regions, or reflects racial or ethnic prejudice. The assessment focuses on language, tone, and representation, aiming to identify harmful biases or imbalanced portrayals.
Gender Bias	GenB	(Guo et al., 2024; Dhamala et al., 2021; Sheng et al., 2021)	T	This criterion evaluates whether the generated content reinforces traditional gender stereotypes or presents biased views based on gender. It examines whether certain roles, traits, or behaviors are unfairly associated with specific genders, such as linking nurturing to women or leadership to men.
Cultural Bias	CulB	(Guo et al., 2024)	T	This criterion evaluates the model's understanding of and respect for cultural diversity. It assesses whether the generated text appropriately acknowledges cultural differences and avoids reinforcing harmful generalizations or stereotypes about specific cultures. For instance, models may misinterpret culturally specific idioms or expressions, or fail to recognize regional language variants, which can result in outputs that lack cultural sensitivity or misrepresent certain communities.
Regional Bias	RegB	(Guo et al., 2024; Sheng et al., 2021)	T	This criterion evaluates whether the generated content shows a bias toward or against any particular nationality or region. It ensures that all national or regional identities are treated fairly, without favoritism, discrimination, or negative stereotyping.
Ethnic Bias	EthB	(Lee & Seong, 2024; Dhamala et al., 2021; Sheng et al., 2021)	T	This criterion assesses whether the generated content reflects racial or ethnic biases, either by perpetuating stereotypes or marginalizing certain groups. It looks for whether certain ethnic or racial groups are portrayed negatively or unfairly advantaged, reinforcing harmful biases.
Toxicity	Tox	(Basile et al., 2019; Chang et al., 2024; Guo et al., 2023)	T	This criterion evaluates whether the model avoids generating harmful, violent, explicit, or inappropriate content. It ensures the system does not produce material that could cause emotional harm, violate legal norms, or infringe on privacy. This includes preventing the generation of discriminatory, NSFW, or harmful language and images.
Harmfulness	Har	(Andriushchenko et al., 2024; Chang et al., 2024; Guo et al., 2023)	T	This criterion evaluates whether the generated content promotes, glorifies, or depicts any form of violence, abuse, or harm—be it physical, emotional, or psychological. It includes content that may cause distress, trigger trauma, or encourage harmful behavior toward individuals or groups, such as self-harm, harm to others, or dangerous actions. Images that cause psychological discomfort when seen are harmful.
Privacy Violation	Priv	(Yao et al., 2024; Chang et al., 2024; Guo et al., 2023)	T	This criterion evaluates whether the generated content reveals personal, private, or sensitive information about individuals without consent, violating privacy norms or laws. This includes disclosing private details such as names, locations, health conditions, or other confidential information. The content should respect an individual's privacy rights, in accordance with legal and ethical standards.
Readability	Rea	(Taylor & Wahlstrom, 1999; Hu et al., 2024a;b; Sugawara et al., 2017)	T	Readability measures how easily a text is understood by its intended audience. It focuses on overall fluency and the clarity of ideas to ensure they are presented simply and align with the readers' comprehension levels. It evaluates the quality of both inter-sentences and intra-sentences, checking if they are grammatically correct, naturally written, with clear meanings, and good context-relatedness and coherence, allowing readers to absorb information effortlessly.
Coherence	Coh	(Liu et al., 2023b; Hu et al., 2024a;b)	T	Coherence refers to the logical consistency and interconnection of ideas within the text. It ensures that the text follows a clear, organized progression, with each section logically linking to the next. A coherent text allows the reader to easily follow the author's argument or narrative, with smooth transitions and no sudden shifts in topic or gaps in reasoning.
Clarity	Cla	(Hu et al., 2024b; Li et al., 2023b)	T	Clarity measures how easily the text can be understood. It emphasizes the use of straightforward, unambiguous language, avoiding complex vocabulary and convoluted sentence structures that could confuse the reader.
Text Repetition	TexR	(Hu et al., 2024b; Wang et al., 2024)	T	Text repetition refers to the unnecessary duplication of phrases, sentences, or ideas that do not add new meaning. Repetition disrupts the flow, reduces clarity, and can make the text feel redundant.
Sentence Complexity	SenC	(Moell & Boye, 2025; Heinz & Idsardi, 2011)	T	This criterion evaluates how sentence structures affect readability. Complex or overly intricate sentences may hinder comprehension, while simpler structures enhance clarity.
Fluency	Flu	(Hu et al., 2024b; Li et al., 2023b; Liu et al., 2023b; Hu et al., 2024a)	T	Fluency measures the smoothness of the language and sentence construction, ensuring the text flows seamlessly and is grammatically correct. The fluent text flows naturally, uses familiar phrasing, and avoids awkward or forced transitions.
Vocabulary Complexity	VocC	(Moell & Boye, 2025; Heinz & Idsardi, 2011)	T	This criterion evaluates whether the vocabulary used in the text is well-suited to the target audience's comprehension level. The language should prioritize clarity by avoiding overly complex or obscure terminology that may hinder understanding, while also ensuring it is not excessively simplistic for the intended readers.
Grammaticality	Gra	(Hu et al., 2024b;a)	T	Grammaticality measures how well the text adheres to the rules of grammar, including sentence structure, verb tense, subject-verb agreement, and punctuation.

Table 14: The definition and source of Universal Aspects (UAs). (Part 2)

Aspect	Abbr	Source	Target	Definition
Spelling Accuracy	SpeA	(Hu et al., 2024b;a)	T	This criterion assesses the correctness of spelling in the text, ensuring there are no typographical errors or misspelled words.
Tense Consistency	TenC	(Hu et al., 2024b;a)	T	This criterion assesses whether the sentence maintains consistent tense usage (past, present, future) throughout. The tense should remain uniform unless a change is contextually required, such as when indicating a change in the time frame.
Structure Accuracy	StrA	(Hu et al., 2024b;a)	T	This criterion evaluates the grammatical accuracy of sentences, focusing on the proper alignment of subject-verb-object and the correct linkage of main and subordinate clauses. It ensures that sentences are constructed following standard grammatical rules, which enhances clarity and readability.
Image Quality	ImaQ	(Liu et al., 2024a; Xu et al., 2023a; 2024)	I	Image quality primarily evaluates the fidelity of the generated image, ensuring it accurately represents the described objects and their characteristics. It also assesses whether the image contains any harmful elements, such as offensive content, misleading representations, or harmful stereotypes.
Fidelity	Fid	(Xu et al., 2023a; 2024)	I	Fidelity evaluates how accurately the generated image represents the shape, characteristics, and behavior of the objects described, ensuring they align with real-world expectations. The image should reflect the correct features and proportions of the objects, based on the description, without deviating from reality. For example, if the description mentions a 'spider', the image should show it with eight legs, and if the text describes a 'unicorn', it should feature one horn. Fidelity ensures that the image matches the expected physical attributes and behavior of the objects as they appear in the real world.
Aesthetic	Aes	(Liao et al., 2025; Xu et al., 2024)	I	Aesthetic measures how pleasing and emotionally striking an image is. It checks if the image combines elements like color and light to create a visually appealing experience that resonates with viewers. This measure helps assess whether the image captures beauty and evokes an emotional response, making it memorable and impactful.
Image Coherence	ImaC	(Liu et al., 2024a)	ITI	Image coherence measures the consistency in style and subject representation across images. This includes textures, color palettes, lighting, and rendering styles, and semantic consistency, which covers consistency of physical attributes, clothing, behavioral traits, and scenes. Image coherence also penalizes image duplication, where the output images are too similar to each other or contain near-duplicate content within the same output.
Image Repetition	ImgR	(Liu et al., 2024a)	ITI	This criterion penalizes the repetition or excessive similarity between images. It identifies cases where multiple images are too similar, either due to identical content, framing, or visual elements, without adding meaningful variety to the overall set.
Stylistic Consistency	StyC	(Liu et al., 2024a)	ITI	This criterion evaluates the uniformity of the artistic style across different images. It requires that all images share the same visual style, such as all realistic photographs, all cartoon illustrations, or all oil paintings. For example, if one image is a cartoon-style illustration, all related images should follow the same cartoon style, rather than mixing different styles like realism, oil painting, or animation.
Semantic Consistency	SemC	(Liu et al., 2024a; Bordalo et al., 2024; Chen et al., 2019)	ITI	This criterion evaluates the consistency of semantic elements in the images across a series. These semantic elements include object attributes, scenes, etc. It ensures that each element within the sequence aligns with the overarching theme or purpose, promoting a clear and unified narrative or instructional flow. For example, a cooking tutorial showing step-by-step cake preparation maintains coherence by following the logical sequence of the recipe, whereas mixing unrelated steps disrupts the flow and confuses the viewer.
Character Consistency	ChaC	(Liu et al., 2024a; Bordalo et al., 2024; Chen et al., 2019)	ITI	This criterion ensures that objects or characters maintain the same attributes (e.g., color, shape, size) across different images. It checks that the physical characteristics of a character or object remain constant. This prevents inconsistencies, such as a character suddenly wearing a different outfit or an object changing in size or appearance without explanation.
Action Consistency	ActC	(Liu et al., 2024a; Bordalo et al., 2024; Chen et al., 2019)	ITI	Action Consistency in image sequences ensures that actions depicted across multiple images follow a logical progression and match in activity and intent. This helps maintain a smooth narrative flow, such as a person going from jogging to running, rather than abruptly switching to an unrelated action like swimming.
Scene Consistency	SceC	(Bordalo et al., 2024; Chen et al., 2019)	ITI	This criterion evaluates how effectively a sequence of images maintains the consistency of the scene elements, such as location and background. For instance, in a photo story about a day at the farmer's market, each image should consistently depict the market environment from dawn to dusk. If a photo suddenly shifts to a beach setting, it disrupts the scene consistency, confusing the viewer and breaking the narrative flow.

Table 15: The definition and source of Universal Aspects (UAs). (Part 3)

Aspect	Abbr	Source	Target	Definition
Text-Image Relationship	T-IR	(Liu et al., 2024a)	ITI	Text-Image Relationship refers to how text and image work together to enhance each other. This includes how the text describes what's shown in the image—objects, their properties, and spatial relationships—and also two additional aspects: Complement, which captures how the text adds extra context or background information beyond mere description to enrich meaning, and Integration, which evaluates how coherently and naturally the text and image are combined into a unified whole. Effective text-image alignment ensures the text not only describes the image but also, through Complement and Integration, enriches its overall meaning.
Alignment	Ali	(Xiong et al., 2024; Huang et al., 2023; Ghosh et al., 2023; Xu et al., 2023a)	ITI	Alignment between text and image is evaluated based on how well the image matches the details in the text. This includes the correct objects, their attributes (e.g., color, size), spatial relationships (e.g., left, right), and actions. The alignment should consider multiple dimensions. For example, if the text describes "two boys wearing white clothes, one on the left and one on the right, playing badminton," the image should show two boys in white clothes, placed correctly in the scene, with a badminton game happening between them.
Object Relationship	ObjR	(Li et al., 2024a; Huang et al., 2023)	ITI	Object Relationship evaluates how accurately the spatial relationships and actions between objects described in the text are represented in the image. It focuses on whether the interactions, positions, and movements of objects in the image match those described in the text. For example, in the description 'a person standing next to a table and reaching for a book', the image should show the person next to the table with their hand extended toward the book. As long as the spatial relationships and actions are correctly represented, the image will be considered accurate.
Spatial Alignment	SpaA	(Gokhale et al., 2022; Ghosh et al., 2023; Li et al., 2024a; Huang et al., 2023)	ITI	This criterion measures how accurately the spatial relationships between objects in the generated output match the descriptions in the text. It evaluates whether the relative positions, orientations (up, down, left, right), and arrangements of objects align with the textual description. For example, if the text mentions a chair to the left of a table, the image should show the chair on the left side of the table, and if a person is sitting on a chair in front of a table, this relationship should also be reflected accurately in the image.
Action Alignment	ActA	(Li et al., 2024a; Huang et al., 2023)	ITI	Action Alignment evaluates how accurately the actions or interactions described in the text are represented in the image. It specifically measures dynamic actions like 'running', 'talking', or 'holding', and does not consider static descriptions or spatial relationships (e.g., left, right, or object arrangements). For example, if the text says 'a dog running after a ball', the image should show the dog running after the ball.
Count Alignment	CouA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	ITI	Count Alignment evaluates how accurately the number of objects described in the text matches both the number and type of objects presented in the corresponding image. It specifically measures whether the quantity of objects mentioned in the text (e.g., two dogs, three cars, or five apples) is correctly reflected in the image, and ensures that the objects shown are of the correct type. For example, if the text mentions 'two dogs', the image should contain two dogs, not two cats. This evaluation focuses solely on the count and type of objects, without considering their attributes or positions.
Object Alignment	ObjA	(Li et al., 2024a; Ghosh et al., 2023)	ITI	Object Alignment evaluates how accurately the objects described in a text are represented in an image. The objects mentioned in the text should be accurately presented in the image, with no omissions. For example, in the description ' A man sitting by himself at the dinner table', the image should feature both a man and a dinner table to achieve a high score. Missing or extra objects in the image would lower the score.
Scene Alignment	ScnA	(Li et al., 2024a)	ITI	Scene Alignment assesses the extent to which the setting or background described in a text is accurately represented in an image. It focuses on elements such as location, weather conditions, time of day, and general atmosphere rather than specific objects. For example, if a text describes 'a sunny environment with historical architecture around a town square', the image should depict this specific setting, showing historical architecture situated within a town square context, not isolated in an open field or desert.
Attribute Alignment	AttA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	ITI	Attribute Alignment assesses how well the attributes (such as color, texture, size, and shape) of objects described in the text are reflected in the image. It ensures that key features like color and size match what's described. For instance, if the text says "a green bicycle," the image should show a bicycle that is green. It also checks if details like texture and shape are accurate, like a "smooth, red ball" being depicted as both smooth and red. Furthermore, the relative size and proportions of objects should be consistent with the description. Overall, the image should visually match the text's description of each object's attributes.
Color Alignment	ColA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	ITI	This criterion assesses whether the colors described in the text match the corresponding objects in the image based solely on the text description, or whether the colors of the objects conform to realistic colors. If the text specifies the color of an object, the evaluation checks whether the image reflects that color. For example, if the text says 'a blue car', the image should show a car that matches the described color, not one of a different color.

Table 16: The definition and source of Universal Aspects (UAs). (Part 4)

Aspect	Abbr	Source	Target	Definition
Shape Alignment	ShaA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	ITI	This criterion evaluates how accurately the shapes described in the text correspond to the shapes of objects in the image, based solely on the text's shape descriptors, or whether the shape of the objects conforms to realistic shape. The text may include common shape terms such as square, round, rectangular, oval, triangular, cubic, cylindrical, spherical, and other basic or specialized shapes like pyramidal, conical, pentagonal, teardrop, crescent, and diamond.
Texture Alignment	TexA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	ITI	This criterion evaluates how accurately the textures described in the text are reflected in the image, based solely on the text description, or whether the textures of the objects conform to realistic textures. It includes surface qualities such as smoothness, roughness, or granularity, as well as material types like wood, plastic, and rubber. Additionally, it may encompass tactile sensations or patterns, such as silky, polished, or other material-based descriptions. The accuracy of the texture description is assessed by comparing the textual description to the actual appearance and surface properties of the objects in the image.
Advanced Alignment	AdvA	(Li et al., 2024a)	ITI	Advanced Alignment is a comprehensive evaluation aspect that assesses the ability to accurately represent complex textual descriptions in an image. It encapsulates Comparison, Negation, and Universality aspects. This means it checks whether the image correctly depicts comparisons between entities based on attributes such as number or volume; represents absence or contradiction of elements indicated by words like 'no' or 'not'; and illustrates scenarios where every member of a group shares a specific attribute or relation.
Comparison	Cmpa	(Li et al., 2024a)	ITI	Comparison evaluates the ability to accurately depict comparisons between entities as described in a text. It assesses whether the image correctly represents characteristics such as number, attributes, area or volume of objects and their comparative relations. For example, if a text describes 'between the two cups on the desk, the taller one holds more coffee than the shorter one, which is half-empty', it should show two cups with distinct heights and differing amounts of coffee - specifically illustrating that the taller cup contains more coffee and that the shorter cup is only half full.
Negation	Nega	(Li et al., 2024a)	ITI	Negation evaluates the ability to accurately depict the absence or contradiction of elements as described in a text. It assesses whether the image correctly represents scenarios where entities are not present, as indicated by words such as 'no', 'not', or 'without'. For example, if a text describes 'a bookshelf with no books, only picture frames', it should show a bookshelf devoid of books but filled with picture frames instead.
Universal	Univ	(Li et al., 2024a)	ITI	Universal evaluates the ability to accurately depict scenarios where every member of a group shares a specific attribute or is involved in a common relation, as described in a text. It assesses whether the image correctly represents situations indicated by words such as 'every', 'all', 'each', or 'both'. For example, if a text describes 'a bustling kitchen where every chef is preparing a dish', it should show all chefs within the kitchen actively engaged in cooking.
Complement	Comp	(Vempala & Preoţiuc-Pietro, 2019)	ITI	This criterion assesses how text and image work together to enhance or clarify the overall meaning. It evaluates how they complement each other, whether through redundancy, emotional amplification, added context, or interpretation, creating a richer understanding than either could provide alone.
Complementary Overlap	ComO	(Vempala & Preoţiuc-Pietro, 2019)	ITI	This criterion captures cases where the image and text redundantly represent the same information, and the image does not contribute anything beyond what is already stated in the text. The text alone is sufficient to convey the core message. For example, a caption saying "Tacos are the best" can be paired with a photo of tacos. Both the image and text communicate the same idea with no additional contextual or semantic contribution from the image.
Emotional Complement	EmoC	(Vempala & Preoţiuc-Pietro, 2019)	ITI	This criterion evaluates how the image, particularly through elements like emojis or memes, enhances or clarifies the emotional tone of the text. The image should convey emotions, moods, or subtle feelings that the text alone might not fully express. This could involve using visual elements such as facial expressions, symbols, or humorous images that directly amplify, complement, or interpret the emotional content of the text, helping to create a more vivid or relatable emotional context.
Comment Complement	ComC	(Vempala & Preoţiuc-Pietro, 2019)	ITI	This criterion assesses how the text offers commentary or evaluation that adds meaning to the image. The text may provide an interpretation, critique, or personal perspective on the image, offering a deeper understanding or a subjective view of what the image represents. The text can either affirm or challenge the implied meaning of the image, creating a more nuanced interaction between the two modalities.
Contextual Complement	ConC	(Vempala & Preoţiuc-Pietro, 2019)	ITI	This criterion evaluates how the text and the image provide complementary background information or contextual details that enhance the understanding of each other. The text should add context, explanations, or relevant background information that helps interpret the image more fully, or conversely, the image should provide visual context that clarifies or enriches the meaning of the text.

Table 17: The definition and source of Task-specific Aspects (TAs). (Part 1)

Aspect	Abbr	Source	Sub-Task	Definition
Coverage	Cov	(Li et al., 2023b)	Summarization	The summary should cover all essential points from the original text, ensuring no significant information is omitted. In this case, completeness is more important than brevity.
Length Constraint	LenC	(Li et al., 2023b)	Summarization	This measures the model's ability to respect any given length restrictions for the response (e.g., sentence count, paragraph count, or specific character limits), ensuring the model does not generate an overly short or excessively long response. If the instruction does not mention a length requirement, all responses should receive a score of -1, without considering the influence of other criteria.
Layout	Lay	(Li et al., 2023b)	Summarization	The written summary is encouraged to follow a clear and well-organized layout. You may use headings, bullet points, lists, tables, or other formatting devices to improve readability. This criterion is not applicable if the user does not explicitly require a specific layout.
Conciseness	Con	(Li et al., 2023b)	Summarization	This criterion evaluates whether the summary is brief and to the point, effectively capturing the essential information of the original content while avoiding unnecessary details or excessive elaboration. It focuses on the ability to convey key elements clearly and concisely, maintaining the core meaning of the text without including irrelevant or redundant content.
Helpfulness	Hel	(Li et al., 2024f; Liu et al., 2023a)	General Visual Conversation	Consider whether the generated text provides valuable insights, additional context, or relevant information that contributes positively to the user's comprehension of the image. Assess whether the model accurately follows any specific instructions or guidelines provided in the prompt. Evaluate the overall contribution of the response to the user experience.
Instruction Following	InsF	(Zeng et al., 2023; Liu et al., 2023a)	General Visual Conversation	Instruction Following is defined as the ability to accurately interpret and execute the task as outlined in the given instruction, focusing solely on providing the correct answer. It measures whether the response directly addresses the task's core requirements, strictly adheres to any specified constraints on response length and format, and does not introduce unnecessary additional background information, explanations, or tone beyond what is requested.
Transparency	Tran	(Zhang et al., 2023b)	Medical VQA	Transparency is a criterion used to assess how clearly and transparently a model explains its reasoning behind the answer to an academic question. It focuses on the model's ability to provide understandable, logical, and well-supported justifications for its responses. This includes breaking down complex ideas, referencing relevant information, and demonstrating a step-by-step thought process. A high level of explainability ensures that users not only receive a correct answer but also understand how the answer was derived, helping to build trust in the model's reasoning and enhancing learning outcomes.
Terminology	Term	(Zhang et al., 2023b)	Medical VQA	Terminology evaluates the model's ability to correctly use and apply medical terminology when responding to queries related to medical image understanding. This includes the accurate usage of terms specific to anatomy, pathology, imaging techniques, or diagnostic processes. For example, if a user asks about a CT scan showing a lung lesion, the model should mention terms like 'pulmonary nodule,' 'radiographic features,' or 'lesion characterization' rather than generic terms like 'spot' or 'image area.' Proper use of medical terms indicates the model's depth of understanding in the field.
Accuracy	Acc†	(Li et al., 2023b)	Rewriting	Accuracy in Rewriting refers to how correctly the changes follow the given instructions. Any modifications are made according to the task requirements.
Pointing Out	PoiO	(Li et al., 2023b)	Rewriting	This criterion evaluates the clarity and transparency of corrections made to the text. Instead of simply returning a corrected version, the correction process must be clearly outlined. This includes pointing out which parts of the text were changed, explaining how the changes were made, and providing a reason for each correction. This approach helps the user understand not only the mistake but also the logic behind the correction, facilitating better learning and comprehension.
Difference	Dif	(Li et al., 2023b)	Rewriting	This criterion evaluates whether the written text shows a clear difference from the original in terms of its appearance or structure. The text should not simply be a direct copy of the original; instead, it should reflect meaningful changes, such as rephrasing or reformatting. The goal is to ensure that the new text is distinct and provides added value, rather than just repeating the same content in the same form.
Creativity	Cre	(Li et al., 2023b)	Creative Writing	Creativity refers to the originality and imagination in the writing. It evaluates whether the content presents fresh ideas, unique expressions, and an inventive approach. Creativity involves not only the ideas but also how the language, structure, and emotions are used. It looks for originality that breaks from the norm, offering something new and engaging for the reader.
Engagingness	Eng	(Li et al., 2023b)	Creative Writing	Engagingness refers to the extent to which a text, whether in creative writing, captures and maintains the audience's interest. This criterion evaluates the ability of the content to stimulate a compelling exchange of ideas, emotions, or information, encouraging active participation and sustained engagement. It measures how well the narrative or dialogue keeps the audience engaged and thought-provoking throughout.

Table 18: The definition and source of Task-specific Aspects (TAs). (Part 2)

Aspect	Abbr	Source	Sub-Task	Definition
Theme	The	(Li et al., 2023b)	Creative Writing	This criterion evaluates whether the response maintains a clear and consistent central theme or subject matter throughout the creative work, such as a song or piece of writing. For example, if the song centers around a romantic breakup, every verse and chorus should consistently reflect and support this theme, creating a cohesive narrative. The goal is to ensure that all elements of the work work together to reinforce the main idea, helping the audience to easily follow and connect with the message being conveyed.
Relevance	Rel	(Li et al., 2023b)	Functional Writing	Relevance measures how closely the content of a written piece aligns with the specific objectives and purpose of the task. It ensures that the text stays on topic, directly addresses the intended subject, and provides information that is appropriate and meaningful for the given context.
Contextual Information	ConI	(Li et al., 2023b)	Functional Writing	This criterion assesses how effectively extra context or elaboration is provided when necessary, enhancing the user's understanding without overwhelming them. It ensures that the information is valuable and directly aligned with the user's inquiry, offering clarity and depth where needed
Example Quality	ExaQ	(Li et al., 2023b)	Functional Writing	This criterion assesses how effectively specific instances or cases are provided to illustrate the concept being discussed. Examples help make abstract ideas more tangible and understandable by showing how they manifest in real-world or hypothetical scenarios.
Accuracy	Acc	(Li et al., 2023b)	General Communication	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
Information Richness	InfoR	(Li et al., 2023b)	General Communication	This criterion assesses whether the response provides sufficient information to address the query, including necessary context and explanations. It ensures that the answer is detailed and clear, offering enough depth to fully address the user's needs without omitting key information.
Understandability	Und	(Li et al., 2023b)	General Communication	This criterion evaluates understandability, which refers to how well a conversational system interprets user inputs and generates contextually relevant responses. It focuses on the model's ability to grasp the meaning of user queries and the flow of the conversation, ensuring responses are appropriate. For example, if a user asks, 'What's the weather like today?', the model should respond with a relevant, clear answer about the current weather, instead of a generic or unrelated reply. A model with high understandability maintains seamless, clear communication.
Suitability	Suit	(Li et al., 2023b)	Title Generation	This criterion evaluates how closely the generated title aligns with the content of the text or work. It's about ensuring that the title directly reflects the main theme or subject matter, providing a meaningful and appropriate summary for the given context. For instance, if the text is about the effects of climate change on wildlife, a relevant title might be Wildlife in Distress: Unraveling the Impact of Climate Change.
Coverage	Cov†	(Li et al., 2023b)	Title Generation	This criterion evaluates how much of the essential information from the source content is captured in the generated title. It's about assessing how well the title encapsulates the key points or main ideas of the source content. For instance, if the text discusses the history, types, and health benefits of meditation, a title that covers these points might be Meditation: Tracing Its Roots and Understanding Its Health Wonders.
Appeal	Appe	(Li et al., 2023b)	Title Generation	This criterion evaluates the attractiveness of the generated title. It assesses whether the title has the power to catch the reader's attention, spark interest, and entice the reader to delve into the content. For example, a title like 'Unlocking Creativity: Harnessing the Power of an Unconventional Mind' might appeal to readers interested in personal development and creativity.
Translation Fidelity	TrnF	(Li et al., 2023b)	Translation	This criterion evaluates the precision of the translation in replicating the source text's content in the target language. It involves correct vocabulary choices, proper grammar, and syntax. It ensures that the meaning, tone, and nuances of the original text are preserved without adding, omitting, or distorting information.
Thoroughness	Thor	(Li et al., 2023b)	Translation	This criterion assesses whether the translation covers all parts of the source text without leaving out any information, details, or nuances. It ensures that the translation fully conveys the original text's content, including all its ideas, arguments, and points, in the target language.
Precision	Prec	Li et al. (2023b)	Data Analysis	This criterion evaluates the precision and correctness of the data analysis results. It assesses the degree to which the results correspond to the true or actual values, ensuring that the output is free from errors or distortions.
Format	Form	(Li et al., 2023b)	Data Analysis	This criterion evaluates the clarity and user-friendliness of the data analysis output generated by the model. It assesses whether the output is well-structured, follows standard data presentation conventions, and is easy to interpret and understand. The output should present the analysis results in a clear and concise manner, using appropriate tables, charts, or other visual aids as necessary. The use of clear headings, labels, and explanations should also be considered.

Table 19: The definition and source of Task-specific Aspects (TAs). (Part 3)

Aspect	Abbr	Source	Sub-Task	Definition
Engagingness	Eng	(Li et al., 2023b)	Question Generation	This criterion evaluates the ability of the generated questions to captivate and hold the attention of the audience. Engaging questions spark curiosity, stimulate thought, and encourage active participation. For instance, a question like 'How does climate change impact our daily lives?' might engage readers because it connects a global issue to personal experiences.
Difficulty	Diff	(Li et al., 2023b)	Question Generation	This criterion assesses whether the complexity level of the generated questions matches the knowledge and comprehension level of the intended audience. For example, for a high school audience, a question like 'Can you explain Einstein's theory of relativity?' might be too difficult, while 'What is the capital of France?' might be too easy.
Relevance	Rel†	(Li et al., 2023b)	Keywords Extraction	In the context of keyword extraction, this criterion evaluates how closely the extracted keywords align with the main themes or subjects of the text. Relevant keywords should accurately reflect the core ideas, topics, or concepts presented in the text, providing a concise and meaningful representation of the content.
Representation	Repr	(Li et al., 2023b)	Keywords Extraction	In the context of keyword extraction, this criterion assesses the extent to which the extracted keywords represent all the significant themes, topics, or concepts in the text. A high coverage means that the keywords collectively provide a comprehensive summary of the text, leaving no important aspect unrepresented.
Accuracy	Acc	(Li et al., 2024f; Zhang et al., 2023c)	Text Rich Understanding	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
Instruction Following	InsF	(Li et al., 2024f; Zeng et al., 2023)	Text Rich Understanding	Instruction Following is defined as the ability to accurately interpret and execute the task as outlined in the given instruction, focusing solely on providing the correct answer. It measures whether the response directly addresses the task's core requirements, strictly adheres to any specified constraints on response length and format, and does not introduce unnecessary additional background information, explanations, or tone beyond what is requested.
Instruction Inconsistency	InsI	(Li et al., 2024f; Huang et al., 2025)	Robustness-oriented Instructions	Instruction inconsistency refers to the model's outputs that deviate from a user's directive. While some deviations might serve safety guidelines, the inconsistencies here signify unintentional misalignment with non-malicious user instructions. For example, the user's actual intention is translation, However, the LLM erroneously deviated from the user's instruction and performed a question-answering task instead.
Context Inconsistency	ConI	(Li et al., 2024f; Huang et al., 2025)	Robustness-oriented Instructions	Context inconsistency occurs when the model's output contradicts the contextual information provided by the user, which may include text or images. This happens when the model misinterprets the context or incorrectly identifies elements, such as attributes of objects in an image or non-existent details. For example, if the user states that the source of the Nile is in the Great Lakes region of central Africa, but the model incorrectly claims it is in the Americas. Similarly, if an image shows a girl in red clothes, but the model mistakenly identifies her as a boy in blue clothes.
Accuracy	Acc	(Li et al., 2023a; 2024f)	Medical Image Understanding	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
Instruction Following	InsF	(Zeng et al., 2023; Masry et al., 2022)	Chart Reasoning	Instruction Following is defined as the ability to accurately interpret and execute the task as outlined in the given instruction, focusing solely on providing the correct answer. It measures whether the response directly addresses the task's core requirements, strictly adheres to any specified constraints on response length and format, and does not introduce unnecessary additional background information, explanations, or tone beyond what is requested.
Accuracy	Acc	(Li et al., 2024f; Masry et al., 2022)	Chart Reasoning	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
Instruction Following	InsF	(Zeng et al., 2023; Mathew et al., 2022)	Graph Reasoning	Instruction Following is defined as the ability to accurately interpret and execute the task as outlined in the given instruction, focusing solely on providing the correct answer. It measures whether the response directly addresses the task's core requirements, strictly adheres to any specified constraints on response length and format, and does not introduce unnecessary additional background information, explanations, or tone beyond what is requested.
Accuracy	Acc	(Li et al., 2024f; Mathew et al., 2022)	Graph Reasoning	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
Reasonableness	Reas	(Liu et al., 2024a)	Storytelling Generation	This criterion evaluates the logical coherence and believability of the generated events within the context of the narrative. It assesses whether the events follow a logical progression, align with the established facts and rules of the narrative world, and are consistent with the characters' motivations and behaviors. The events should not introduce any inconsistencies or contradictions that could disrupt the audience's suspension of disbelief. Furthermore, the events should be plausible and realistic, avoiding any absurd or improbable scenarios unless they are justified by the narrative context.

Table 20: The definition and source of Task-specific Aspects (TAs). (Part 4)

Aspect	Abbr	Source	Sub-Task	Definition
Instruction Following	InsF	(Zeng et al., 2023; Singh et al., 2019)	Text Reasoning	Instruction Following is defined as the ability to accurately interpret and execute the task as outlined in the given instruction, focusing solely on providing the correct answer. It measures whether the response directly addresses the task's core requirements, strictly adheres to any specified constraints on response length and format, and does not introduce unnecessary additional background information, explanations, or tone beyond what is requested.
Accuracy	Acc	(Li et al., 2024f; Singh et al., 2019)	Text Reasoning	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
Conciseness	Con†	(Mokady et al., 2021; Hossain et al., 2019)	Simple Image Captioning	This criterion examines whether the caption remains concise and focused, conveying the necessary information efficiently without including redundant or overly detailed content. It emphasizes clarity and precision in communicating key points while minimizing distractions from irrelevant elements.
Object Alignment	ObjA	(Mokady et al., 2021; Hossain et al., 2019)	Simple Image Captioning	Object Alignment evaluates the precision with which images match the objects described in the corresponding text. This criterion ensures that the objects mentioned in the text are accurately represented in the image, focusing solely on whether the object itself is present. It does not consider additional criteria such as attribute alignment.
Text-Image Relationship	T-IR	(Mokady et al., 2021; Hossain et al., 2019)	Simple Image Captioning	Text-Image Relationship refers to how text and image work together to enhance each other. This includes how the text describes what's shown in the image—objects, their properties, and spatial relationships—and also two additional aspects: Complement, which captures how the text adds extra context or background information beyond mere description to enrich meaning, and Integration, which evaluates how coherently and naturally the text and image are combined into a unified whole. Effective text-image alignment ensures the text not only describes the image but also, through Complement and Integration, enriches its overall meaning.
Alignment	Ali	(Mokady et al., 2021; Hossain et al., 2019)	Simple Image Captioning	Alignment between text and images ensures the image matches the description, including the attributes and relationships of objects. For example, if the text is 'a cat dressed as Napoleon Bonaparte,' the image must show a cat. If the text is 'a little girl sitting in front of a sewing machine,' the image should show only one girl in that setting, not multiple girls. Misalignment happens when the subject is missing or the image doesn't match the description.
Completeness	Com	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	Completeness evaluates whether the caption covers all significant elements in the image, including key objects, actions, and possible contextual information (e.g., background, setting, or scene details). A complete caption should mention all essential features of the image without omitting critical information.
Attribute Alignment	AttA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	Attribute Alignment assesses how well the attributes (such as color, texture, size, and shape) of objects described in the text are reflected in the image. It ensures that key features like color and size match what's described. For instance, if the text says "a green bicycle," the image should show a bicycle that is green. It also checks if details like texture and shape are accurate, like a "smooth, red ball" being depicted as both smooth and red. Furthermore, the relative size and proportions of objects should be consistent with the description. Overall, the image should visually match the text's description of each object's attributes.
Spatial Alignment	SpaA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	In the image captioning generation task, this criterion measures how accurately the spatial relationships between objects in the text align with their representation in the image. It evaluates whether the relative positions and arrangements of objects described in the text are correctly reflected in the image, focusing solely on spatial relationships. It does not consider other criteria, such as object attributes or additional context.
Count Alignment	CouA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	Count Alignment evaluates how accurately the number of objects described in the text matches both the number and type of objects presented in the corresponding image. It specifically measures whether the quantity of objects mentioned in the text (e.g., two dogs, three cars, or five apples) is correctly reflected in the image, and ensures that the objects shown are of the correct type. For example, if the text mentions 'two dogs', the image should contain two dogs, not two cats. This evaluation focuses solely on the count and type of objects, without considering their attributes or positions.
Action Alignment	ActA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	In the image captioning generation task, Action Alignment evaluates how accurately the actions or interactions depicted in the image are described in the text. It specifically measures actions like 'running', 'talking', or 'holding', and does not take into account static descriptions or spatial relationships (e.g., left, right, or object arrangements). For example, if the image shows a dog running after a ball, the text should include the word 'running'.
Object Alignment	ObjA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	Object Alignment evaluates the precision with which images match the objects described in the corresponding text. This criterion ensures that the objects mentioned in the text are accurately represented in the image, focusing solely on whether the object itself is present. It does not consider additional criteria such as attribute alignment.

Table 21: The definition and source of Task-specific Aspects (TAs). (Part 5)

Aspect	Abbr	Source	Sub-Task	Definition
Color Alignment	ColA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	In the image captioning generation task, this metric assesses whether the colors described in the text match the corresponding objects in the image. If the image presents an object with a specific color, the text should describe it as accurately as possible. For example, if the image shows a blue car and a person wearing a yellow shirt, the text should mention 'blue' and 'yellow'. The more accurately the colors are described, the higher the score.
Texture Alignment	TexA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	In the image captioning generation task, this metric checks how accurately the textures presented in the image are described in the text. If the objects in the image have specific texture qualities, such as surface qualities (e.g., smooth, rough), material types (e.g., wooden, metallic), or tactile sensations or patterns (e.g., silky, polished), the more accurately these textures are expressed in the text, the higher the score.
Shape Alignment	ShaA	(Hossain et al., 2019; Zhao et al., 2023)	Detailed Image Captioning	In the image captioning generation task, this metric assesses how accurately the shapes described in the text match the shapes of objects in the image. If the image contains objects with shape descriptors such as geometric terms (e.g., cubic, circular, triangular) or more general attributes like symmetry, curvature, complexity, and flatness, the text should include corresponding descriptions. The more accurately the shapes are described, the higher the score.
Expertise	Expe	(Li et al., 2023c)	Academic QA	Expertise refers to the model's ability to accurately use domain-specific terminology when answering academic questions. This includes correctly applying terms relevant to the subject, such as those in science, history, or literature. For example, when asked about a historical event, the model should use terms like 'primary sources' or 'historical context' instead of vague words like 'thin' or 'event'. Proper use of academic terms shows the model's understanding of the subject.
Correctness	Corr	(Li et al., 2023c)	Academic VQA	The conciseness metric assesses whether the model provides clear, direct, and succinct answers to academic questions, avoiding unnecessary details. The accuracy metric evaluates if the responses correctly address the query based on the provided academic context, ensuring precise and relevant information is delivered.
Explainability	Expl	(Li et al., 2023c)	Academic VQA	Explainability is a criterion used to assess how clearly a model explains its reasoning when answering an academic question. It focuses on whether the model provides clear, logical steps and justifications for its answer, making it easy for users to understand how the answer was reached.
Interpretability	Intp	(Chen et al., 2023)	Embodied Decision-making	Interpretability in the context of embodied decision-making refers to how well a model can articulate its decisions and actions. It's about transparency, where the model should be able to provide clear explanations for why it chose a particular course of action over others. For example, if an AI playing chess makes a move, interpretability would involve the AI being able to explain why it made that specific move, considering factors like board position, potential threats, and future opportunities.
Completion	Comp	(Chen et al., 2023)	Embodied Decision-making	Task completion is essentially about goal achievement. In embodied decision-making tasks, this means whether or not the model has successfully navigated its environment to achieve its intended objective. This could range from simple objectives, such as reaching a certain location within an environment (like an AI navigating through a maze), to more complex ones, like completing multiple interrelated tasks in sequence (like cooking, where you need to chop vegetables before frying them).
Riskiness	Risk	(Chen et al., 2023)	Embodied Decision-making	"Riskiness measures how well the model avoids harmful situations or risky behaviors during task execution. This is particularly important in real-world applications where mistakes can have serious consequences. For instance, if we consider autonomous driving again, safety here doesn't just mean avoiding collisions; it also involves obeying traffic rules, not taking unnecessary risks while overtaking other vehicles, and maintaining safe distances from pedestrians and other road users.
Originality	Orig	(Liu et al., 2024a)	Activitynet Continuation	This criterion evaluates the originality and inventiveness in the generation of a mixed media fairytale. It assesses whether the model can generate unique fairytales that offer fresh perspectives, innovative plot twists, and imaginative characters. The creativity also extends to the integration of visuals and text, with the model expected to generate images that not only complement the narrative but also enhance it in unexpected and creative ways. The model should be able to break away from common fairytale tropes and cliches, offering a unique and memorable narrative experience.
Simplification	Simp	(Liu et al., 2024a)	Activitynet Continuation	This criterion evaluates the suitability of the generated fairytale for a young audience. It assesses whether the model can generate a narrative that is easy to understand and digest for children, without compromising the richness and depth of the story. The language used should be simple, clear, and age-appropriate, avoiding complex vocabulary or convoluted sentence structures. The plot should be straightforward and easy to follow, with no overly complicated or confusing elements. The visuals should also be simple and clear, effectively conveying the events and emotions of the story in a way that is accessible and engaging for children. This criterion ensures that the generated fairytale is not only enjoyable but also appropriate and beneficial for its intended young audience.

Table 22: The definition and source of Task-specific Aspects (TAs). (Part 6)

Aspect	Abbr	Source	Sub-Task	Definition
Fulfillment	Full	(Liu et al., 2024a)	Storytelling Generation	This criterion evaluates whether the generated events are comprehensive and well-rounded within the narrative. It assesses whether each event has a clear beginning, middle, and end, providing a full and satisfying narrative arc. The events should not feel fragmented or incomplete, but should provide all the necessary details and developments to understand and appreciate their significance in the narrative. This includes the introduction of the event, the unfolding of the event, and the resolution or outcome of the event.
Alignment	Ali	(Xiong et al., 2024; Huang et al., 2023; Ghosh et al., 2023; Xu et al., 2023a)	Text-to-Image Generation	Alignment between text and image is evaluated based on how well the image matches the details in the text. This includes the correct objects, their attributes (e.g., color, size), spatial relationships (e.g., left, right), and actions. The alignment should consider multiple dimensions. For example, if the text describes "two boys wearing white clothes, one on the left and one on the right, playing badminton," the image should show two boys in white clothes, placed correctly in the scene, with a badminton game happening between them.
Object Relationship	ObjR	(Li et al., 2024a; Huang et al., 2023)	Text-to-Image Generation	Object Relationship evaluates how accurately the spatial relationships and actions between objects described in the text are represented in the image. It focuses on whether the interactions, positions, and movements of objects in the image match those described in the text. For example, in the description 'a person standing next to a table and reaching for a book', the image should show the person next to the table with their hand extended toward the book. As long as the spatial relationships and actions are correctly represented, the image will be considered accurate.
Spatial Alignment	SpaA†	(Ghosh et al., 2023; Li et al., 2024a; Huang et al., 2023; Gokhale et al., 2022)	Text-to-Image Generation	This criterion measures how accurately the spatial relationships between objects in the generated output match the descriptions in the text. It evaluates whether the relative positions, orientations (up, down, left, right), and arrangements of objects align with the textual description. For example, if the text mentions a chair to the left of a table, the image should show the chair on the left side of the table, and if a person is sitting on a chair in front of a table, this relationship should also be reflected accurately in the image.
Action Alignment	ActA†	(Li et al., 2024a; Huang et al., 2023)	Text-to-Image Generation	Action Alignment evaluates how accurately the actions or interactions described in the text are represented in the image. It specifically measures dynamic actions like 'running', 'talking', or 'holding', and does not consider static descriptions or spatial relationships (e.g., left, right, or object arrangements). For example, if the text says 'a dog running after a ball', the image should show the dog running after the ball.
Count Alignment	CouA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	Text-to-Image Generation	Count Alignment evaluates how accurately the number of objects described in the text matches both the number and type of objects presented in the corresponding image. It specifically measures whether the quantity of objects mentioned in the text (e.g., two dogs, three cars, or five apples) is correctly reflected in the image, and ensures that the objects shown are of the correct type. For example, if the text mentions 'two dogs', the image should contain two dogs, not two cats. This evaluation focuses solely on the count and type of objects, without considering their attributes or positions.
Object Alignment	ObjA†	(Li et al., 2024a; Ghosh et al., 2023)	Text-to-Image Generation	Object Alignment evaluates how accurately the objects described in a text are represented in an image. The objects mentioned in the text should be accurately presented in the image, with no omissions. For example, in the description ' A man sitting by himself at the dinner table', the image should feature both a man and a dinner table to achieve a high score. Missing or extra objects in the image would lower the score.
Attribute Alignment	AttA	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	Text-to-Image Generation	Attribute Alignment assesses how well the attributes (such as color, texture, size, and shape) of objects described in the text are reflected in the image. It ensures that key features like color and size match what's described. For instance, if the text says "a green bicycle," the image should show a bicycle that is green. It also checks if details like texture and shape are accurate, like a "smooth, red ball" being depicted as both smooth and red. Furthermore, the relative size and proportions of objects should be consistent with the description. Overall, the image should visually match the text's description of each object's attributes.
Color Alignment	ColA†	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	Text-to-Image Generation	This criterion assesses whether the colors described in the text match the corresponding objects in the image based solely on the text description, or whether the colors of the objects conform to realistic colors. If the text specifies the color of an object, the evaluation checks whether the image reflects that color. For example, if the text says 'a blue car', the image should show a car that matches the described color, not one of a different color.
Completeness	Com*	(Liu et al., 2024a; Koupaee & Wang, 2018)	Multimodal Script Generation	Completeness measures whether the model-generated steps provide a complete guide to finishing the task, covering all necessary actions needed from start to finish.
Feasibility	Fea	(Liu et al., 2024a; Koupaee & Wang, 2018)	Multimodal Script Generation	This criterion evaluates the feasibility of content output by a model for a specific task by assessing whether each step is factually correct, logically sequenced, and detailed enough for successful execution. It ensures that the steps are not only accurate but also practical, providing clear and actionable instructions that lead to the desired outcome without leading to confusion or errors.
Helpfulness	Hel†	(Liu et al., 2024a; Koupaee & Wang, 2018)	Multimodal Script Generation	This criterion evaluates whether the model's response provides clear and practical steps to help the user complete the task, such as planting a strawberry. It checks if the response includes all essential details, like preparing the soil, planting the seeds, and caring for the plant, along with useful tips or warnings.

Table 23: The definition and source of Task-specific Aspects (TAs). (Part 7)

Aspect	Abbr	Source	Sub-Task	Definition
Safety	Saf	(Liu et al., 2024a; Koupaee & Wang, 2018)	Multimodal Script Generation	This criterion checks if the model's responses are safe and avoid harmful content. It ensures the model doesn't suggest dangerous, unethical, or illegal actions, and follows safety and ethical guidelines. The model should also provide safety warnings or advice when needed, prioritizing user well-being and avoiding any risks.
Shape Alignment	ShaA†	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	Text-to-Image Generation	This criterion evaluates how accurately the shapes described in the text correspond to the shapes of objects in the image, based solely on the text's shape descriptors, or whether the shape of the objects conforms to realistic shape. The text may include common shape terms such as square, round, rectangular, oval, triangular, cubic, cylindrical, spherical, and other basic or specialized shapes like pyramidal, conical, pentagonal, teardrop, crescent, and diamond.
Texture Alignment	TexA†	(Li et al., 2024a; Huang et al., 2023; Ghosh et al., 2023)	Text-to-Image Generation	This criterion evaluates how accurately the textures described in the text are reflected in the image, based solely on the text description, or whether the textures of the objects conform to realistic textures. It includes surface qualities such as smoothness, roughness, or granularity, as well as material types like wood, plastic, and rubber. Additionally, it may encompass tactile sensations or patterns, such as silky, polished, or other material-based descriptions. The accuracy of the texture description is assessed by comparing the textual description to the actual appearance and surface properties of the objects in the image.
Advanced Alignment	AdaA	(Li et al., 2024a)	Text-to-Image Generation	Advanced Alignment is a comprehensive evaluation aspect that assesses the ability to accurately represent complex textual descriptions in an image. It encapsulates Comparison, Negation, and Universality aspects. This means it checks whether the image correctly depicts comparisons between entities based on attributes such as number or volume; represents absence or contradiction of elements indicated by words like 'no' or 'not'; and illustrates scenarios where every member of a group shares a specific attribute or relation.
Comparison	Cmpa	(Li et al., 2024a)	Text-to-Image Generation	Comparison evaluates the ability to accurately depict comparisons between entities as described in a text. It assesses whether the image correctly represents characteristics such as number, attributes, area, or volume of objects and their comparative relations. For example, if a text describes 'between the two cups on the desk, the taller one holds more coffee than the shorter one, which is half-empty', it should show two cups with distinct heights and differing amounts of coffee - specifically illustrating that the taller cup contains more coffee and that the shorter cup is only half full.
Negation	Nega	(Li et al., 2024a)	Text-to-Image Generation	Negation evaluates the ability to accurately depict the absence or contradiction of elements as described in a text. It assesses whether the image correctly represents scenarios where entities are not present, as indicated by words such as 'no', 'not', or 'without'. For example, if a text describes 'a bookshelf with no books, only picture frames', it should show a bookshelf devoid of books but filled with picture frames instead.
Universal	Univ	(Li et al., 2024a)	Text-to-Image Generation	Universal evaluates the ability to accurately depict scenarios where every member of a group shares a specific attribute or is involved in a common relation, as described in a text. It assesses whether the image correctly represents situations indicated by words such as 'every', 'all', 'each', or 'both'. For example, if a text describes 'a bustling kitchen where every chef is preparing a dish', it should show all chefs within the kitchen actively engaged in cooking.
Accuracy	Acc*	(Zhang et al., 2023a)	Image Editing	Accuracy measures how well the edited image aligns with the given instructions in the text. The model should accurately implement the specified changes, including adjusting colors, shapes, positions, and adding or removing elements as instructed. This includes adhering to any quantity or specific attribute requirements described in the text. For example, if the text requests 'add one small white dog', the image should show exactly one small white dog. If the model generates two white dogs or a black dog instead, the accuracy evaluation should be poor.
Creativity	Cre†	(Smilevski et al., 2018; Liu et al., 2024a)	Visual Story Completion	Creativity in visual story completion refers to the originality and imagination in developing the narrative. It evaluates how fresh ideas, unique expressions, and an inventive approach are used to expand upon the visual elements. Creativity looks for originality in the interpretation of visuals, offering a new and engaging continuation of the story that captivates the reader.
Engagingness	Eng†	(Smilevski et al., 2018; Liu et al., 2024a)	Visual Story Completion	Engagingness refers to the degree to which a story fosters continued interaction, maintains or elevates interest, and stimulates a compelling exchange of ideas, emotions, or information. This criterion assesses the ability of the content to engage the audience, encouraging active participation and sustained engagement by making the narrative or dialogue both interesting and thought-provoking.
Completeness	Com†	(Smilevski et al., 2018; Liu et al., 2024a)	Visual Story Completion	Completeness measures whether the model-generated visual story provides a complete narrative, covering all necessary elements (e.g., characters, settings, actions, and events) from the beginning to the end of the story. For example, if the input is a sequence of images showing a person preparing for a hike, the generated story should include all key steps such as packing a backpack, driving to the trail, starting the hike, and reaching the destination, ensuring a coherent and comprehensive narrative.

Table 24: Aspects used by evaluators in the benchmark evaluations. (Part 1)

	Subclass	Aspect	Definition
SummEval	Coherence	Coherence	Coherence refers to the logical consistency and interconnection of ideas within the text. It ensures that the text follows a clear, organized progression, with each section logically linking to the next. A coherent text allows the reader to easily follow the author's argument or narrative, with smooth transitions and no sudden shifts in topic or gaps in reasoning.
	Consistency	Consistency	Evaluate whether the facts presented in the response align with the facts described in the article. This metric focuses solely on ensuring the accuracy of the events or facts, without considering the level of detail. As long as the facts in the response match those in the article, both the response and the article should receive the same score. The evaluation does not include assessing whether key information is omitted.
	Consistency	Coverage	The summary should cover all essential points from the original text, ensuring no significant information is omitted. In this case, completeness is more important than brevity.
	Fluency	Fluency	Fluency measures the smoothness and naturalness of language and sentence construction, ensuring the text flows seamlessly and is grammatically correct. Fluent text not only maintains a natural flow but also includes concise phrasing, where brief answers can still achieve high scores, as long as they remain fluent.
MANS	-	Relevance	Relevance in story generation measures how closely the generated story aligns with the specific requirements, title, or provided initial part of the story. It ensures that the narrative stays on topic, directly addresses and develops from the given prompt or initial storyline, and provides a continuation that is logical and easy to understand within the established context.
MT-Bench	-	Correctness	This criterion evaluates the model's ability to generate responses that not only align with the specific instructions provided but also enhance the user's understanding of the context. It checks if the model adheres to any given guidelines about format and word count, while also providing correct and insightful answers. A high-quality response should offer valuable insights, additional context, or relevant information that contributes positively to the user's comprehension. This assessment ensures the model's output is not just accurate and instruction-compliant, but also informative, contextually rich, and user-friendly.
WildVision	-	Overall	This criterion is a comprehensive assessment of the model's responses. It combines both accuracy and helpfulness. Accuracy is a measure that ensures the model's responses are factually correct and in alignment with the user's expectations, while helpfulness evaluates the value and relevance of the information in the responses, focusing on how much they enhance the user's understanding. An overall evaluation takes into account how correctly the model interprets and responds to the user's prompts, as well as how beneficial the responses are to the user. It's a holistic measure of the model's effectiveness in providing accurate, insightful, and useful responses.

Table 25: Aspects used by evaluators in the benchmark evaluations. (Part 2)

	Subclass	Aspect	Definition
MLLM-as-a-Judge	COCO	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
	C.C.	Alignment	Alignment between text and image is evaluated based on how well the details in the text correspond to the visual elements in the image. This includes the accurate representation of objects, their attributes (e.g., color, size), spatial relationships (e.g., positioning, direction), and actions. The alignment should ensure that the text captures all relevant details from the image. For example, if the image shows a surfer riding a wave with the surfboard visibly moving, the text should accurately describe these elements, such as the wave's size, the surfer's stance, the motion of the surfboard, and the wave's characteristics, ensuring all critical visual details are reflected in the caption.
	Diff	Alignment Accuracy Analysis	This criterion evaluates the model's ability to accurately analyze the alignment between a given text and image. It assesses if the model's responses correctly identify the quality of text-image pairing, pinpointing perfect alignment or errors in object attributes and spatial relationships. For instance, if the text mentions a cat, but the image shows a dog, responses that accurately point out this misalignment in objects are considered high quality. The ability to accurately identify issues or provide precise evaluations determines the quality of the responses.
	Graphics	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
	Math	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
	Text	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
	WIT	Alignment Accuracy Analysis	This criterion evaluates the model's ability to accurately analyze the alignment and relevance between a given text and image. It assesses if the model's responses correctly identify the quality of text-image pairing, pinpointing perfect alignment or errors in object attributes and spatial relationships. For instance, if the text mentions a cat, but the image shows a dog, responses that accurately point out this misalignment in objects are considered high quality. The ability to accurately identify issues or provide precise evaluations determines the quality of the responses.
	Chart	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.

Table 26: Aspects used by evaluators in the mata-benchmark evaluations. (Part 3)

	Subclass	Aspect	Definition
MLLM-as-a-Judge	CC-3M	Helpfulness	This criterion evaluates the quality and relevance of the model's response. It focuses on whether the information provided is not only accurate but also of significant value to the user. The evaluation takes into account if the model's output offers beneficial insights that can help deepen the user's understanding of the topic at hand. Additionally, it assesses if the model can provide further context or pertinent information that can enhance the user's comprehension or perspective. The goal is to ensure that the model's response is not merely factual, but also insightful and enriching, contributing positively to the user's knowledge and understanding.
	VisIT	Helpfulness	This criterion evaluates the quality and relevance of the model's response. It focuses on whether the information provided is not only accurate but also of significant value to the user. The evaluation takes into account if the model's output offers beneficial insights that can help deepen the user's understanding of the topic at hand. Additionally, it assesses if the model can provide further context or pertinent information that can enhance the user's comprehension or perspective. The goal is to ensure that the model's response is not merely factual, but also insightful and enriching, contributing positively to the user's knowledge and understanding.
	SciQA	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.
	M2W	Helpfulness	This criterion evaluates the quality and relevance of the model's response. It focuses on whether the information provided is not only accurate but also of significant value to the user. The evaluation takes into account if the model's output offers beneficial insights that can help deepen the user's understanding of the topic at hand. Additionally, it assesses if the model can provide further context or pertinent information that can enhance the user's comprehension or perspective. The goal is to ensure that the model's response is not merely factual, but also insightful and enriching, contributing positively to the user's knowledge and understanding.
	Aes	Aesthetic	This criterion evaluates the model's response in terms of its accuracy and depth in explaining the aesthetic quality of the image. It assesses whether the model correctly identifies and articulates the aesthetic elements present in the image, such as color harmony, composition, lighting, and texture. Furthermore, it examines if the model offers a comprehensive analysis that considers multiple aspects of aesthetics, including but not limited to symmetry, balance, contrast, and the rule of thirds. The model's ability to relate these elements to the overall aesthetic appeal of the image is also taken into account.
	MM-Vet	Precision	This criterion evaluates the accuracy of numerical calculations in a model's response when answering questions that require both image recognition and computation. It focuses on whether the model correctly identifies relevant visual elements from the image and performs accurate calculations to derive the final answer. The score is based solely on the correctness of the numerical result, ensuring the model's ability to handle tasks that combine visual analysis and quantitative reasoning effectively.

Table 27: Aspects used by evaluators in the benchmark evaluations. (Part 4)

	Subclass	Aspect	Definition
GenAI-Bench & Pick-a-Pic	-	Alignment	Alignment between text and image is evaluated based on how well the image matches the details in the text. This includes the correct objects, their attributes (e.g., color, size), spatial relationships (e.g., left, right), and actions. The alignment should consider multiple dimensions. For example, if the text describes "two boys wearing white clothes, one on the left and one on the right, playing badminton," the image should show two boys in white clothes, placed correctly in the scene, with a badminton game happening between them.
Winoground	Relation	Contextual Relationship	This criterion checks if a model correctly shows what's described in a prompt, like the number of people, their actions, or their emotions. It pays close attention to whether the roles and actions in the image precisely match those specified. For example, if the prompt says 'Two humans and one wheel', it ensures the image has two people and one wheel, not the other way around. Similarly, if it states 'the taller person hugs the shorter person', the taller individual should be the one doing the hugging. It's all about ensuring that the picture matches the words exactly, both in terms of quantity and the specifics of the actions.
	Object	Alignment	Alignment between text and image is evaluated based on how well the image matches the details in the text. This includes the correct objects, their attributes (e.g., color, size), spatial relationships (e.g., left, right), and actions. The alignment should consider multiple dimensions. For example, if the text describes "two boys wearing white clothes, one on the left and one on the right, playing badminton," the image should show two boys in white clothes, placed correctly in the scene, with a badminton game happening between them.
	Both	Alignment	Alignment between text and image is evaluated based on how well the image matches the details in the text. This includes the correct objects, their attributes (e.g., color, size), spatial relationships (e.g., left, right), and actions. The alignment should consider multiple dimensions. For example, if the text describes "two boys wearing white clothes, one on the left and one on the right, playing badminton," the image should show two boys in white clothes, placed correctly in the scene, with a badminton game happening between them.
VLRewardBench	General	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. In cases where two responses are provided, only one must be correct, and the other may contain errors or contradictions.
	Hullucination	Factual Contradiction	Factual Contradiction occurs when the caption mentions something that is not present in the image, which indicates a hallucination. If a caption refers to an object or scene that cannot be found in the image, it means the model has generated incorrect or imaginary content. In this context, one of the two responses will always contain a hallucination, where the mentioned content cannot be found in the image. Captions that accurately describe only what is visible in the image, without inventing anything, will score higher.
	Reasoning	Accuracy	Accuracy measures how correct the model's response is, ensuring it is both factually accurate and aligned with the user's expectations. It evaluates the model's ability to interpret the user's question correctly and provide a response based on verified objective facts. The output must not only be correct but also reflect reliable and accurate information.

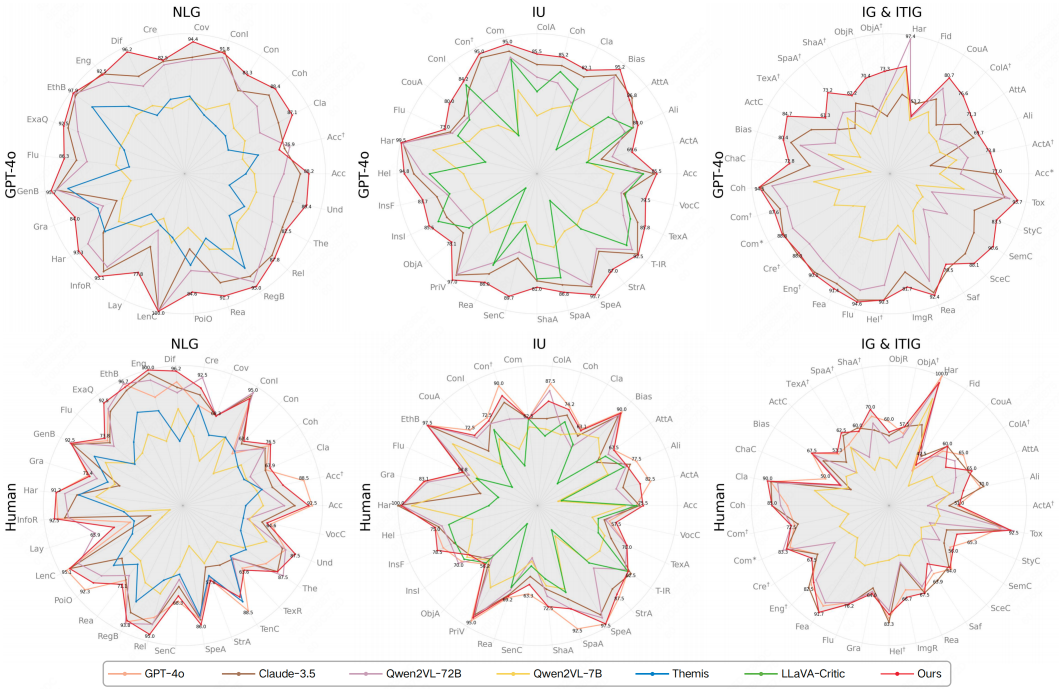


Figure 12: Comparison on FRA-ID and FRA-ID-H. The top shows alignment with GPT-4o using FRA-ID, the bottom shows alignment with human annotators using FRA-ID-H. Each point is the average accuracy (with ties) for an aspect shared across sub-tasks of the same task.

Table 28: Average accuracy on FRA-ID and FRA-ID-H, separated by UAs and TAs. Average accuracy (with ties) is calculated by averaging the evaluation accuracies across sub-tasks within each task.

Method	Seen UAs for Seen Tasks								Seen TAs for Seen Tasks							
	FRA-ID (GPT-4o)				FRA-ID-H (Human)				FRA-ID (GPT-4o)				FRA-ID-H (Human)			
	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG
GPT-4o	-	-	-	-	**76.1**	**79.1**	65.0	66.1	-	-	-	-	79.5	**74.8**	**58.8**	73.6
Claude-3.5	79.4	79.1	55.2	77.2	67.2	71.9	52.5	61.1	80.3	78.0	52.7	86.4	76.3	64.8	57.8	**75.9**
Qwen2VL-72B	78.1	74.0	**70.0**	58.5	70.1	70.5	**67.5**	56.9	76.1	68.8	53.4	75.4	79.1	67.4	47.8	70.3
Qwen2VL-7B	51.1	50.0	58.4	37.4	48.6	55.6	57.7	38.2	48.9	46.9	42.6	39.5	46.7	51.6	34.6	36.1
Themis	60.4	-	-	-	60.3	-	-	-	47.5	-	-	-	53.4	-	-	-
LLaVA-Critic	-	47.5	-	-	-	39.0	-	-	-	68.0	-	-	-	56.0	-	-
Ours	**91.1**	**88.2**	61.0	**87.0**	75.8	78.2	65.0	**70.0**	**87.9**	**85.1**	**71.9**	**87.2**	**84.1**	71.5	58.1	73.9

D More Experiment Results

D.1 In-domain Evaluation

For in-domain evaluation, we assess performance using the FRA-ID and FRA-ID-H. Figure 12 presents radar charts that illustrate the overall degree of alignment. The accompanying Table 28 reports average accuracies for each task.

The results show that UFEval achieves higher alignment with both GPT-4o and human annotators compared to the baselines (see shaded area in radar chart). Specifically, excluding GPT-4o, UFEval surpasses all baselines on NLG, IU, and ITIG tasks within FRA-ID and FRA-ID-H, achieving overall average accuracies of 82.4% and 72.1%, respectively. On NLG, UFEval outperforms Themis, while our base model Qwen2VL-7B performs much worse. Moreover, UFEval shows strong performance on the ITIG of FRA-ID-H, where its alignment with human annotators exceeds that of GPT-4o. This improvement can be attributed to UFEval's generalization ability and cross-task transfer performance.

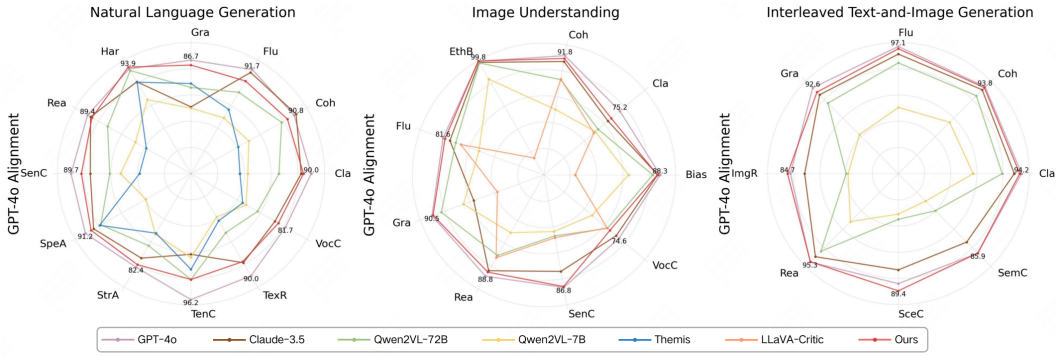


Figure 13: Comparison on FRAUAs-OOD to evaluate the generalization in UAs.

D.2 Experimental Results for FRAUAs-OOD

In the training set, for each sub-task, according to our aspect selection procedure, if the task output modality includes text, all aspects under the Text Branch in the UAs Tree can be selected. However, once the model learns a UAs in one sub-task, it can leverage this UAs for inference in other sub-tasks without requiring retraining. Additionally, to maintain sample balance between UAs and TAs during training (as selecting all aspects under the Text Branch for a text-output task would lead to an imbalance in the number of UAs and TAs in the training set), we selectively sample a subset of corresponding UAs in the training set.

To validate that UFEval can evaluate UAs in sub-tasks where these UAs were not encountered during training, we construct FRAUAs-OOD, which comprises the same sub-tasks as the training set but introduces unseen UAs across 4 tasks. The experimental results are shown in Figure 13 and Table 29. The figure illustrates that UFEval exhibits robust generalization capability despite not being trained on the complete set of UAs for each sub-task.

Table 29: The results of average accuracy on FRAUAs-OOD for UFEval.

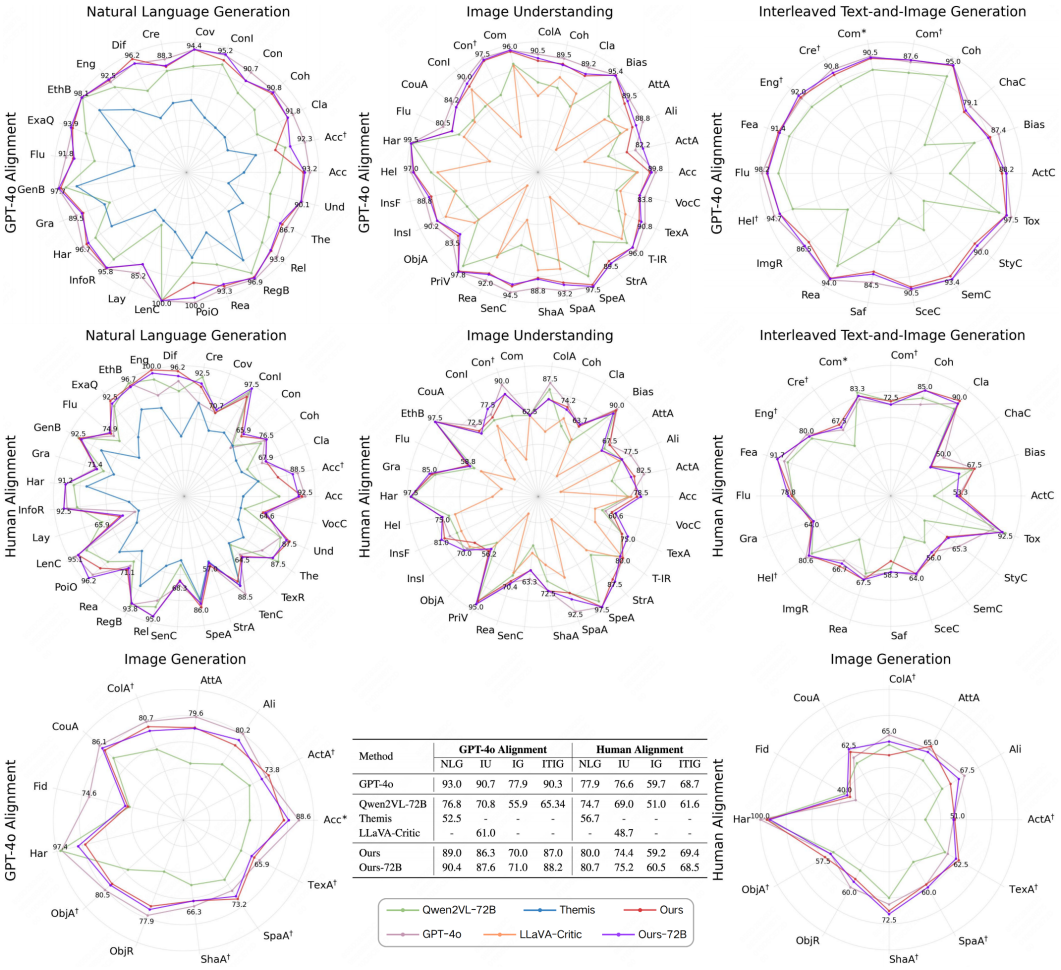
Method	GPT-4o Alignment		
	NLG	IU	ITIG
Claude-3.5	77.3	77.6	83.0
Qwen2VL-72B	69.7	71.4	65.3
Qwen2VL-7B	49.9	56.5	44.4
Themis	54.2	-	-
LLaVA-Critic	-	50.0	-
Ours	**82.8**	**83.1**	**90.4**

D.3 Experiment Results of UFEval-72B

We also train UFEval-72B using the same fine-tuning method as UFEval-7B with Qwen2-VL-72B-Instruct as the backbone and test it under the same experiment settings. The results, shown in Figures 14 and 15 and Tables 30 to 32, indicate that UFEval-72B offers some improvement over UFEval-7B.

Table 30: Evaluation as MLLM-as-a-Judge for IG task. We evaluate baselines across three benchmarks.

Method	GenAI-Bench		Winoground				Pick-a-Pic	
	tau(↑)	diff(↑)	Relation diff (↑)	Object diff (↑)	Both diff (↑)	Ave. diff (↑)	tau(↑)	diff(↑)
GPT-4o	**55.6**	69.5	62.6	**73.0**	73.0	69.5	**54.4**	**59.2**
Claude-3.5	**55.6**	**71.0**	**71.2**	**73.0**	69.2	**71.1**	49.1	53.7
Qwen2VL-72B	49.1	52.6	46.7	60.9	57.6	55.0	38.6	38.3
Qwen2VL-7B	35.8	38.0	33.4	42.5	46.1	40.6	38.1	40.2
VisionReward	51.0	66.4	60.2	64.1	74.9	66.4	48.9	58.0
ImageReward	48.6	64.9	54.0	58.2	69.2	60.4	48.8	55.7
Ours	53.6	65.5	57.5	59.1	**80.7**	65.7	50.0	57.3
Ours-72B	55.5	69.0	61.5	63.4	62.5	62.4	52.0	58.6

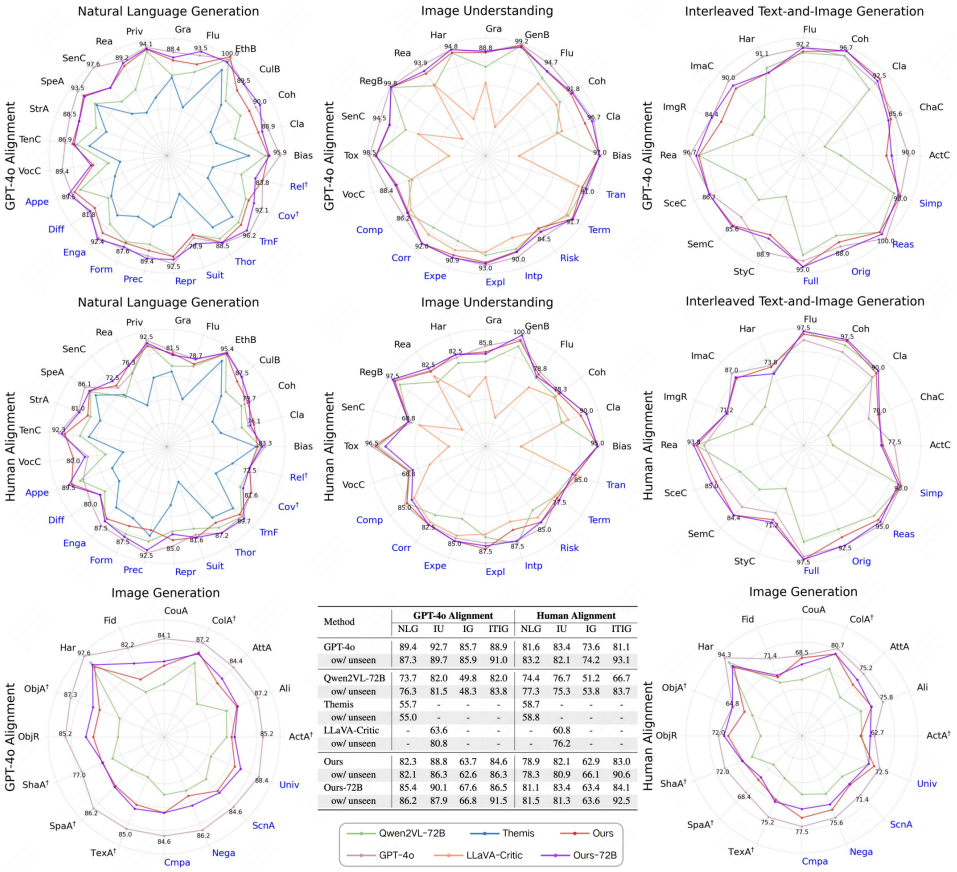


Method	GPT-4o Alignment				Human Alignment			
	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG
GPT-4o	93.0	90.7	77.9	90.3	77.9	76.6	59.7	68.7
Qwen2VL-72B	76.8	70.8	55.9	65.34	74.7	69.0	51.0	61.6
Themis	52.5	-	-	-	56.7	-	-	-
LLaVA-Critic	-	61.0	-	-	-	48.7	-	-
Ours	89.0	86.3	70.0	87.0	80.0	74.4	59.2	69.4
Ours-72B	90.4	87.6	71.0	88.2	80.7	75.2	60.5	68.5

Figure 14: Comparison of UFEval-72B on FRA-ID and FRA-ID-H. The top shows alignment with GPT-4o using FRA-ID, the bottom shows alignment with human annotators using FRA-ID-H. Each point is the average accuracy (with ties) for an aspect shared across sub-tasks of the same task.

Table 31: Evaluation as MLLM-as-a-Judge for IU tasks. We evaluate baselines across three benchmarks.

Method	WildVision			MLLM-as-a-Judge		VLRewardBench			
	tau (↑)	diff (↑)	τ (↑)	tau (↑)	diff (↑)	General diff (↑)	Hallucination diff (↑)	Reasoning diff (↑)	Ave. diff (↑)
GPT-4o	**55.3**	**70.1**	**73.3**	58.1	67.0	<u>50.2</u>	<u>81.4</u>	**74.8**	**68.8**
Claude-3.5	53.3	67.3	61.2	**58.4**	<u>68.3</u>	38.5	**82.6**	66.1	62.4
Qwen2VL-72B	50.3	59.6	65.5	54.6	58.5	**50.8**	75.4	70.7	<u>65.6</u>
Qwen2VL-7B	39.2	40.6	23.1	41.3	44.6	45.1	62.8	62.5	56.8
LLaVA-Critic	53.0	66.0	59.6	55.6	65.5	42.0	41.2	60.0	47.7
Ours	53.9	<u>68.6</u>	66.5	57.2	67.0	46.4	57.7	71.1	58.4
Ours-72B	<u>54.5</u>	<u>68.6</u>	<u>69.9</u>	<u>58.2</u>	**73.1**	44.3	61.6	<u>72.1</u>	59.3



Method	GPT-4o Alignment				Human Alignment			
	NLG	IU	IG	ITIG	NLG	IU	IG	ITIG
GPT-4o	89.4	92.7	85.7	88.9	81.6	83.4	73.6	81.1
ow/ unseen	87.3	89.7	85.9	91.0	83.2	82.1	74.2	93.1
Qwen2VL-72B	73.7	82.0	49.8	82.0	74.4	76.7	51.2	66.7
ow/ unseen	76.3	81.5	48.3	83.8	77.3	75.3	53.8	83.7
Themis	55.7	-	-	-	58.7	-	-	-
ow/ unseen	55.0	-	-	-	58.8	-	-	-
LLaVA-Critic	-	63.6	-	-	-	60.8	-	-
ow/ unseen	-	80.8	-	-	-	76.2	-	-
Ours	82.3	88.8	63.7	84.6	78.9	82.1	62.9	83.0
ow/ unseen	82.1	86.3	62.6	86.3	78.3	80.9	66.1	90.6
Ours-72B	85.4	90.1	67.6	86.5	81.1	83.4	63.4	84.1
ow/ unseen	86.2	87.9	66.8	91.5	81.5	81.3	63.6	92.5

Figure 15: Comparison of UFEval-72B on FRA-OOD and FRA-OOD-H. The blue-colored aspects indicate unseen TAs, whereas the black-colored aspects represent seen UAs. The 'ow/ unseen' designation in the table represents evaluations conducted exclusively on unseen aspects, with accuracy metrics computed only for unseen TAs.

Table 32: Evaluation as MLLM-as-a-Judge for NLG task.

Method	SummEval										MANS	MT-Bench	
	Coherence		Consistency		Fluency		Relevance		Ave.				
	tau(↑)	diff(↑)	tau(↑)	diff(↑)	tau(↑)	diff(↑)	tau(↑)	diff(↑)	tau(↑)	diff(↑)	diff(↑)	tau(↑)	diff(↑)
GPT-4o	58.0	64.2	79.1	85.1	64.3	72.8	60.1	67.1	65.3	72.3	68.5	70.9	83.5
Claude-3.5	63.5	70.6	**81.6**	**87.9**	73.7	83.1	59.6	66.6	**69.6**	77.0	68.4	76.3	90.7
Qwen2VL-72B	**66.8**	**73.2**	66.8	66.4	71.8	80.4	**62.2**	**69.3**	66.9	72.3	19.3	75.9	88.7
Qwen2VL-7B	48.8	52.5	35.5	30.6	44.8	49.8	41.7	44.5	42.7	44.3	60.2	44.5	50.1
Themis	60.7	62.1	81.8	86.0	73.3	77.7	54.4	54.6	67.5	70.1	44.2	43.6	37.7
Auto-J	-	-	-	-	-	-	-	-	60.4	67.0	68.2	73.0	85.5
Prometheus 2	55.2	62.1	65.5	74.7	61.6	69.8	55.0	61.8	59.3	67.1	69.0	55.1	72.0
Ours	64.6	71.8	75.2	83.5	74.7	84.5	61.3	67.6	69.0	76.9	**69.3**	74.9	88.3
Ours-72B	64.6	71.7	77.2	86.9	**75.1**	**84.7**	60.3	67.0	69.3	**77.6**	68.0	**77.4**	**91.1**

D.4 Specific Experimental Results

We report the specific average accuracy of each evaluator for every aspect within each task in Figure 12 and Figure 3 of the main text. The alignment is measured against: (1) GPT-4o on the FRA-ID and FRA-OOD, and (2) human annotators on the FRA-ID-H and FRA-OOD-H. Detailed results are presented in Tables 33 to 48. The reported average accuracy is computed by aggregating the same aspect across different sub-tasks within the same task. We also include the comprehensive results on MLLM-as-a-Judge in Table 49. The specific accuracy of multi-aspect assessment learning is show in Table 50 and 52.

Table 33: The specific experimental results of the IG task in the FRA-ID are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Accuracy	Acc*	88.6	75.6	50.8	35.9	77.0
Action Alignment	ActA†	71.4	29.8	57.1	48.8	73.8
Alignment	Ali	80.2	65.3	52.0	39.1	69.7
Attribute Alignment	AttA	79.6	61.8	43.8	40.7	71.3
Color Alignment	ColA†	80.7	48.3	57.9	40.7	76.6
Count Alignment	CouA	86.1	62.8	71.9	52.1	80.7
Fidelity	Fid	74.6	53.2	42.7	40.8	44.0
Harmfulness	Har	96.5	57.4	97.4	76.1	77.7
Object Alignment	ObjA†	80.5	41.2	57.1	53.5	73.3
Object Relationship	ObjR	77.9	49.7	42.0	38.9	70.4
Shape Alignment	ShaA†	66.3	56.1	50.0	32.0	62.2
Spatial Alignment	SpaA†	65.5	41.2	55.3	50.4	73.2
Texture Alignment	TexA†	65.9	48.5	49.8	37.4	61.3

Table 34: The specific experimental results of the ITIG task in the FRA-ID are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Consistency	ActC	88.2	64.9	45.2	24.5	84.7
Bias	Bias	87.4	74.7	67.6	46.2	80.4
Character Consistency	ChaC	79.1	58.2	33.7	14.2	72.8
Coherence	Coh	95.0	93.7	85.3	55.1	94.5
Completeness	Com†	86.0	83.2	77.7	38.2	87.6
Completeness	Com*	90.5	88.0	80.1	46.0	88.8
Creativity	Cre†	90.8	88.0	76.6	35.1	86.5
Engagingness	Eng†	88.8	88.7	78.8	29.6	90.0
Feasibility	Fea	90.7	87.3	82.5	50.2	91.4
Fluency	Flu	98.2	93.1	85.8	49.5	94.6
Helpfulness	Hel†	94.7	90.3	79.6	46.6	90.3
Image Repetition	ImgR	86.5	71.0	43.1	37.3	81.7
Readability	Rea	94.0	90.4	81.5	53.1	92.4
Safety	Saf	84.5	79.5	53.0	31.3	76.1
Scene Consistency	SceC	90.5	78.6	35.0	28.6	88.1
Semantic Consistency	SemC	93.4	73.6	50.0	28.7	90.6
Stylistic Consistency	StyC	90.0	66.1	32.7	20.8	83.5
Toxicity	Tox	97.5	84.2	87.8	54.1	93.7

Table 35: The specific experimental results of the IG task in the FRA-ID-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Alignment	ActA†	50.0	43.8	28.1	34.4	51.0
Alignment	Ali	67.5	70.0	47.5	45.0	55.0
Attribute Alignment	AttA	62.5	62.5	52.5	37.5	65.0
Color Alignment	ColA†	65.0	60.0	57.5	37.5	49.5
Count Alignment	CouA	55.0	60.0	50.0	20.0	60.0
Fidelity	Fid	30.0	42.5	40.0	33.0	35.0
Harmfulness	Har	100	62.5	95.0	82.5	95.0
Object Alignment	ObjA†	52.5	57.5	50.0	27.5	57.5
Object Relationship	ObjR	60.0	50.0	45.0	32.5	52.5
Shape Alignment	ShaA†	65.0	55.0	60.0	35.0	70.0
Spatial Alignment	SpaA†	57.5	60.0	37.5	27.5	57.5
Texture Alignment	TexA†	52.5	60.0	50.0	50.0	62.5

Table 36: The specific experimental results of the NLG task in the FRA-ID are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Themis	Ours
Accuracy	Acc	93.2	82.1	70.3	49.4	43.3	88.2
Accuracy	Acc†	92.3	73.1	76.9	42.3	53.8	69.2
Clarity	Cla	91.8	75.9	60.8	48.8	35.3	87.1
Coherence	Coh	90.8	82.0	70.4	47.8	39.6	88.4
Conciseness	Con	90.7	70.4	72.2	59.3	40.7	83.3
Contextual Information	ConI	94.6	91.8	87.1	48.3	44.2	90.5
Coverage	Cov	94.4	83.3	81.5	53.7	55.6	94.4
Creativity	Cre	88.3	81.7	79.2	47.5	54.2	82.5
Difference	Dif	92.3	76.9	69.2	53.8	46.2	96.2
Engagingness	Eng	90.0	92.5	85.0	54.2	62.5	91.7
Ethnic Bias	EthB	98.1	95.7	97.0	49.6	81.3	97.9
Example Quality	ExaQ	92.5	89.1	81.6	49.7	46.3	92.5
Fluency	Flu	91.8	77.6	70.0	49.3	39.8	86.3
Gender Bias	GenB	97.7	92.1	91.9	38.7	83.6	95.7
Grammaticality	Gra	89.5	52	62.5	47.7	66.5	84.0
Harmfulness	Har	96.7	71.7	85.0	58.0	70.0	93.3
Information Richness	InfoR	95.8	90.5	86.3	49.0	37.3	95.1
Layout	Lay	85.2	57.4	44.4	35.2	40.7	77.8
Length Constraint	LenC	100	100	100	50.0	46.3	100
Pointing Out	PoiO	100	53.8	69.2	38.5	65.4	84.6
Readability	Rea	93.3	81.7	74.2	56.7	48.3	91.7
Regional Bias	RegB	96.9	86.7	91.4	63.4	79.6	95.0
Relevance	Rel	93.9	85.0	78.2	51.0	43.5	87.8
Theme	The	86.7	79.2	70.8	51.7	48.3	82.5
Understandability	Und	90.1	79.1	65.4	52.5	32.7	89.4

Table 37: The specific experimental results of the IU task in the FRA-ID are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	LLaVA-Critic	Ours
Accuracy	Acc	89.8	83.6	75.3	54.5	75.9	85.5
Action Alignment	ActA	76.3	47.2	55.2	37.4	36.4	69.6
Alignment	Ali	88.7	76.2	69.5	42.2	75.5	80.0
Attribute Alignment	AttA	89.5	86.8	69.0	51.7	65.0	84.5
Bias	Bias	95.4	90.2	88.4	49.0	22.1	95.2
Clarity	Cla	89.2	76.8	56.2	53.8	67.1	82.1
Coherence	Coh	89.5	79.2	66.5	47.7	74.8	85.3
Color Alignment	ColA	90.5	78.0	69.0	41.5	58.8	85.5
Completeness	Com	96.0	89.7	85.3	61.3	83.8	95.0
Conciseness	Con†	97.5	92.0	65.7	31.8	30.2	95.0
Context Inconsistency	ConI	90.0	77.5	64.7	44.3	81.2	84.3
Count Alignment	CouA	84.3	66.5	63.7	50.7	61.8	80.0
Fluency	Flu	80.5	68.8	67.0	53.7	62.7	73.0
Harmfulness	Har	99.5	98.1	98.5	76.1	37.8	99.5
Helpfulness	Hel	97.0	88.0	81.8	52.5	76.7	94.8
Instruction Following	InsF	88.8	73.4	65.9	48.4	64.0	83.7
Instruction Inconsistency	InsI	90.2	71.5	60.3	42.9	78.2	85.5
Object Alignment	ObjA	83.5	67.5	58.0	46.1	62.0	78.1
Privacy Violations	PriV	97.5	92.5	91.7	53.5	13.0	97.0
Readability	Rea	92.0	79.2	70.3	39.3	72.3	86.0
Sentence Complexity	SenC	94.5	61.8	57.3	35.0	37.4	89.7
Shape Alignment	ShaA	88.7	76.5	62.5	49.7	75.0	81.0
Spatial Alignment	SpaA	93.2	81.2	69.0	53.7	76.0	86.8
Spelling Accuracy	SpeA	97.5	89.4	88.4	45.4	21.2	95.7
Structure Accuracy	StrA	89.3	70.8	68.3	46.1	50.3	87.0
Text-Image Relationship	T-IR	96.0	91.7	86.8	42.0	81.8	92.5
Texture Alignment	TexA	90.7	79.7	68.5	47.2	74.0	85.8
Vocabulary Complexity	VocC	83.8	63.5	61.6	51.1	64.7	79.5

Table 38: The specific experimental results of the ITIG task in the FRA-ID-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Consistency	ActC	52.0	44.0	33.3	33.3	53.3
Bias	Bias	67.5	57.5	65.0	37.5	67.5
Character Consistency	ChaC	50.0	30.0	37.5	32.5	38.5
Clarity	Cla	87.5	87.5	82.5	55.0	90.0
Coherence	Coh	73.8	85.0	78.7	41.2	85.0
Completeness	Com†	70.0	67.5	72.5	30.0	72.5
Completeness	Com*	77.8	80.6	83.3	38.9	80.6
Creativity	Cre†	62.5	62.5	47.5	35.0	67.5
Engagingness	Eng†	80.0	82.5	75.0	32.5	77.5
Feasibility	Fea	86.1	91.7	83.3	44.4	91.7
Fluency	Flu	75.0	76.2	75.0	46.3	76.0
Grammaticality	Gra	62.5	60.0	62.5	41.2	64.0
Helpfulness	Hel†	80.6	83.3	75.0	38.9	77.8
Image Repetition	ImgR	61.3	42.7	41.3	36.0	66.7
Readability	Rea	65.0	67.5	62.5	38.8	67.5
Safety	Saf	58.3	63.9	55.6	33.3	50.0
Scene Consistency	SceC	64.0	61.3	33.3	33.3	62.7
Semantic Consistency	SemC	53.3	48.0	44.0	28.0	56.0
Stylistic Consistency	StyC	65.3	45.3	32.0	36.0	53.3
Toxicity	Tox	82.5	90.0	92.5	37.5	90.0

Table 39: The specific experimental results of the NLG task in the FRA-ID-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Themis	Ours
Accuracy	Acc	92.5	80.0	72.5	57.5	45	90.0
Accuracy	Acc†	88.5	65.4	57.7	53.8	57.7	73.1
Clarity	Cla	64.7	64.2	52.4	49.2	47.6	67.9
Coherence	Coh	74.9	71.7	71.1	50.8	48.1	76.5
Conciseness	Con	51.2	61.0	53.7	31.7	53.7	63.4
Contextual Information	ConI	92.5	92.5	95.0	57.5	57.5	90.0
Coverage	Cov	68.3	68.3	68.3	34.1	53.7	68.3
Creativity	Cre	72.5	80.0	92.5	50.0	72.5	85.0
Difference	Dif	88.5	84.6	80.8	69.2	46.2	96.2
Engagingness	Eng	80.0	97.5	92.5	47.5	70.0	100
Ethnic Bias	EthB	96.7	91.7	96.7	56.7	74.2	94.2
Example Quality	ExaQ	87.5	85.0	87.5	47.5	57.5	92.5
Fluency	Flu	71.1	69.0	65.2	46.5	45.5	73.8
Gender Bias	GenB	90.8	90.0	92.5	63.3	72.5	91.7
Grammaticality	Gra	71.4	47.6	64.6	51.0	56.5	70.1
Harmfulness	Har	91.2	76.2	85.0	62.5	75.0	91.2
Information Richness	InfoR	90.0	85.0	82.5	40.0	42.5	92.5
Layout	Lay	39.5	24.4	65.9	9.8	36.6	51.7
Length Constraint	LenC	95.1	92.7	85.4	17.1	41.5	92.7
Pointing Out	PoiO	92.3	57.7	76.9	50.0	65.4	84.6
Readability	Rea	67.4	66.8	63.1	49.7	52.9	71.1
Regional Bias	RegB	92.5	85.0	93.8	61.3	76.2	91.2
Relevance	Rel	82.5	87.5	87.5	65.0	55.0	95.0
Sentence Complexity	SenC	68.3	57.8	52.8	43.5	49.1	64.6
Spelling Accuracy	SpeA	82.2	82.2	81.3	29.0	79.4	86.0
Structure Accuracy	StrA	56.1	46.7	54.2	43.9	53.3	57.0
Tense Consistency	TenC	88.5	50.0	80.8	38.5	80.8	76.9
Text Repetition	TexR	61.7	52.3	51.4	35.5	49.5	63.6
Theme	The	72.5	82.5	87.5	52.5	55.0	82.5
Understandability	Und	80.0	77.5	80.0	65.0	45.0	87.5
Vocabulary Complexity	VocC	64.6	57.8	47.2	48.4	44.1	61.5

Table 40: The specific experimental results of the ITIG task in the FRA-OOD are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Consistency	ActC	90.0	62.5	32.2	42.2	71.1
Character Consistency	ChaC	85.6	71.3	20.0	24.4	81.1
Clarity	Cla	87.5	87.5	80.0	50.0	92.5
Coherence	Coh	92.2	95.0	92.2	68.9	96.7
Fluency	Flu	92.2	87.5	87.8	60.0	88.9
Harmfulness	Har	91.1	65.0	81.1	61.1	76.7
Image Coherence	ImaC	90.0	67.1	46.7	30.0	81.1
Image Repetition	ImgR	84.4	75.0	46.7	46.7	75.6
Readability	Rea	96.7	86.3	88.9	57.8	88.9
Scene Consistency	SceC	85.6	76.2	53.3	31.1	86.7
Semantic Consistency	SemC	74.4	73.4	53.3	32.2	84.4
Stylistic Consistency	StyC	88.9	61.3	37.8	16.7	73.3
Fulfillment	Full	90.0	95.0	85.0	42.5	95.0
Originality	Orig	84.0	70.0	74.0	50.0	80.0
Reasonableness	Reas	100	97.5	92.5	30.0	92.5
Simplification	Simp	90.0	77.5	84.0	52.0	90.0

Table 41: The specific experimental results of the IU task in the FRA-ID-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	LLaVA-Critic	Ours
Accuracy	Acc	70.5	73.0	68.5	62.5	71.0	75.5
Action Alignment	ActA	82.5	62.5	57.5	15.0	17.5	75.0
Alignment	Ali	77.5	72.5	69.5	55.0	67.5	70.0
Attribute Alignment	AttA	67.5	52.5	62.5	55.5	60.0	67.5
Bias	Bias	90.0	86.3	87.6	65.0	21.2	88.7
Clarity	Cla	63.1	57.5	48.1	59.4	50.6	61.9
Coherence	Coh	74.2	67.9	56.7	53.3	62.9	72.1
Color Alignment	ColA	87.5	62.5	82.5	55.0	50.0	75.0
Completeness	Com	62.5	62.5	62.5	56.5	62.5	62.5
Conciseness	Con†	90.0	77.5	65.0	42.5	30.0	82.5
Context Inconsistency	ConI	72.5	67.5	67.5	55.0	62.5	70.0
Count Alignment	CouA	72.5	52.5	65.0	40.0	37.5	67.5
Ethnic Bias	EthB	97.5	97.5	97.5	75.0	22.5	97.5
Fluency	Flu	58.3	57.9	53.7	45.8	47.9	58.8
Grammaticality	Gra	82.5	41.9	80.6	57.5	28.7	83.1
Harmfulness	Har	97.5	100	96.8	90.0	35.0	97.5
Helpfulness	Hel	75.0	72.5	70.0	47.5	65.0	75.0
Instruction Following	InsF	74.0	65.5	68.0	51.5	64.5	78.5
Instruction Inconsistency	InsI	67.5	60.0	70.0	62.5	65.0	60.0
Object Alignment	ObjA	56.2	51.2	47.5	52.5	55.0	56.2
Privacy Violations	PriV	95.0	90.0	92.5	67.5	15.0	92.5
Readability	Rea	69.2	67.1	63.3	52.5	61.3	67.9
Sentence Complexity	SenC	63.3	50.8	37.5	40.0	43.3	56.7
Shape Alignment	ShaA	72.5	60.0	65.0	57.5	52.5	72.5
Spatial Alignment	SpaA	92.5	70.0	77.5	62.5	65.0	77.5
Spelling Accuracy	SpeA	97.5	90.0	92.5	25.0	20.0	97.5
Structure Accuracy	StrA	87.5	80.0	60.0	52.5	55.0	85.0
Text-Image Relationship	T-IR	80.0	82.5	77.5	55.0	80.0	80.0
Texture Alignment	TexA	70.0	57.5	70.0	52.5	47.5	70.0
Vocabulary Complexity	VocC	52.5	48.7	50.0	40.0	44.4	57.5

Table 42: The specific experimental results of the ITIG task in the FRA-OOD-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Consistency	ActC	77.5	63.4	25.0	50.0	67.5
Character Consistency	ChaC	60.0	68.7	22.5	35.0	68.8
Clarity	Cla	82.5	90.0	90.0	47.5	87.5
Coherence	Coh	86.3	95.0	92.5	62.5	96.3
Fluency	Flu	90.0	95.0	97.5	65.0	95.0
Harmfulness	Har	73.8	71.3	67.5	65.0	72.5
Image Coherence	ImaC	87.0	67.1	46.7	30.0	81.1
Image Repetition	ImgR	70.0	53.6	47.5	45.0	71.3
Readability	Rea	85.0	90.0	91.2	60.0	91.2
Scene Consistency	SceC	76.2	85.5	58.8	38.8	81.2
Semantic Consistency	SemC	74.4	73.4	53.3	32.2	84.4
Stylistic Consistency	StyC	62.5	75.7	40.0	20.0	68.8
Fulfillment	Full	95.0	97.3	82.5	40.0	97.5
Originality	Orig	92.5	92.9	77.5	55.0	85.0
Reasonableness	Reas	95.0	94.7	85.0	32.5	90.0
Simplification	Simp	90.0	79.3	90.0	50.0	90.0

Table 43: The specific experimental results of the IG task in the FRA-OOD are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Alignment	ActA†	85.2	59.8	54.0	44.4	58.2
Alignment	Ali	87.2	64.8	58.2	39.2	67.8
Attribute Alignment	AttA	84.4	68.8	40.0	44.0	65.0
Color Alignment	ColA†	87.2	59.8	68.6	45.4	78.2
Count Alignment	CouA	84.1	62.8	45.4	36.9	61.1
Fidelity	Fid	82.2	59.4	41.6	43.0	53.2
Harmfulness	Har	97.6	85.6	90.0	79.8	84.8
Object Alignment	ObjA†	87.3	64.0	41.6	26.5	60.2
Object Relationship	ObjR	85.2	62.6	39.7	27.5	61.0
Shape Alignment	ShaA†	77.0	47.0	48.2	44.0	57.8
Spatial Alignment	SpaA†	86.2	59.0	39.6	31.0	59.2
Texture Alignment	TexA†	85.0	70.0	37.0	41.6	62.6
Comparison	Cmpa	84.6	75.9	49.4	33.3	64.8
Negation	Nega	86.2	57.3	49.6	34.0	54.5
Scene Alignment	ScnA	84.6	65.4	51.0	48.8	65.2
Universal	Univ	88.4	63.9	43.5	34.7	66.0

Table 44: The specific experimental results of the IG task in the FRA-OOD-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Ours
Action Alignment	ActA†	62.7	56.6	51.8	42.2	50.6
Alignment	Ali	75.8	61.2	55.1	41.0	65.2
Attribute Alignment	AttA	75.2	62.1	46.6	49.7	62.1
Color Alignment	ColA†	80.7	68.3	62.1	47.2	75.8
Count Alignment	CouA	68.5	74.1	48.1	27.8	66.7
Fidelity	Fid	71.4	51.1	49.3	44.9	54.2
Harmfulness	Har	94.3	80.6	89.0	77.5	82.8
Object Alignment	ObjA†	64.8	68.5	46.3	27.8	53.7
Object Relationship	ObjR	72.0	62.2	35.7	26.6	63.6
Shape Alignment	ShaA†	72.0	50.3	46.6	48.4	55.3
Spatial Alignment	SpaA†	68.4	54.1	34.7	34.7	53.1
Texture Alignment	TexA†	75.2	65.2	39.1	43.5	59
Comparison	Cmpa	77.5	80.0	50.0	37.5	70
Negation	Nega	75.6	61.0	53.7	46.3	68.3
Scene Alignment	ScnA	71.4	57.1	49.1	58.0	58.9
Universal	Univ	72.5	62.5	62.5	42.5	67.5

Table 45: The specific experimental results of the NLG task in the FRA-OOD are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Themis	Ours
Bias	Bias	95.9	64.1	83.5	51.8	69.4	85.9
Clarity	Cla	88.9	74.5	59.3	56.5	43.3	80.0
Coherence	Coh	86.0	80.4	71.1	54.0	42.4	79.6
Cultural Bias	CulB	89.5	52.6	68.4	31.6	63.2	81.6
Ethnic Bias	EthB	97.4	97.4	97.4	68.4	86.8	100
Fluency	Flu	90.0	79.1	75.1	52.4	43.5	82.0
Grammaticality	Gra	88.4	64.5	68.7	47.9	66.8	81.6
Privacy Violation	Priv	94.1	90.0	90.6	64.1	50.6	92.4
Readability	Rea	89.2	74.3	62.5	53.8	40.6	84.7
Sentence Complexity	SenC	97.6	75.3	60.0	57.1	46.5	75.3
Spelling Accuracy	SpeA	93.5	82.8	75.6	30.5	73.7	87.4
Structure Accuracy	StrA	88.5	78.2	64.5	59.5	55.3	80.2
Tense Consistency	TenC	86.9	80.8	76.9	46.2	66.2	80.8
Vocabulary Complexity	VocC	89.4	66.5	50.0	46.5	40.0	62.9
Appeal	Appe	86.8	89.5	78.9	44.7	47.4	84.2
Difficulty	Diff	81.8	76.5	61.4	58.3	58.3	75.8
Engagement	Enga	90.9	86.4	75.8	47.7	65.9	85.6
Format	Form	87.6	84.7	81.2	61.8	57.4	82.4
Precision	Prec	87.1	88.2	77.1	58.8	61.5	82.4
Representation	Repr	92.5	83.8	86.3	57.5	52.5	86.3
Suitability	Suit	73.7	76.3	71.1	60.5	34.2	71.1
Thoroughness	Thor	88.5	88.5	83.8	53.8	72.3	87.7
Translation Fidelity	TrnF	96.2	93.8	83.1	59.2	76.2	86.2
Coverage	Cov†	92.1	76.3	76.3	42.1	39.5	78.9
Relevance	Rel†	83.8	81.2	65.0	55.0	40.0	83.8

Table 46: The specific experimental results of the IU task in the FRA-OOD are used to evaluate alignment with GPT-4o.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	LLaVA-Critic	Ours
Bias	Bias	96.5	93.8	97.0	69.0	28.5	97.0
Clarity	Cla	96.7	80.5	63.5	63.0	68.1	88.7
Coherence	Coh	91.8	87.5	77.9	59.4	74.1	89.5
Fluency	Flu	94.7	86.1	81.6	58.4	75.8	88.5
Gender Bias	GenB	99.3	98.6	98.5	83.8	26.0	98.5
Grammaticality	Gra	88.4	64.4	75.5	63.1	61.9	87.7
Harmfulness	Har	94.8	90.7	91.5	71.8	33.0	92.2
Readability	Rea	93.9	86.8	73.5	62.2	75.0	86.3
Regional Bias	RegB	99.8	100	98.8	76.7	24.5	99.0
Sentence Complexity	SenC	94.5	75.5	71.2	53.4	60.2	85.9
Toxicity	Tox	98.5	93.6	94.8	72.0	31.2	93.2
Vocabulary Complexity	VocC	88.4	72.5	65.6	58.2	68.5	79.3
Completion	Comp	86.3	78.7	82.5	69.0	78.2	81.2
Correctness	Corr	92.0	87.3	75.1	78.5	82.5	92.0
Expertise	Expe	90.7	83.8	75.9	73.6	84.3	88.6
Explainability	Expl	93.0	86.3	84.5	74.9	81.8	90.4
Interpretability	Intp	90.0	82.8	83.0	71.0	73.0	85.3
Riskiness	Risk	84.5	78.0	76.7	72.0	79.0	76.0
Terminology	Term	90.5	87.3	86.4	58.5	85.0	91.0
Transparency	Tran	91.0	91.0	88.2	55.8	82.8	86.2

Table 47: The specific experimental results of the NLG task in the FRA-OOD-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	Themis	Ours
Bias	Bias	83.3	62.8	79.5	53.8	75.6	76.9
Clarity	Cla	71.1	72.6	65.5	51.3	47.2	76.1
Coherence	Coh	75.1	80.7	72.1	39.1	44.2	78.2
Cultural Bias	CulB	87.5	52.6	68.4	31.6	63.2	81.6
Ethnic Bias	EthB	95.4	93.7	93.7	68.4	86.8	94.3
Fluency	Flu	78.2	81.2	72.1	48.2	42.6	74.1
Grammaticality	Gra	81.5	66.2	68.8	50.3	64.3	79.6
Privacy Violation	Priv	92.5	87.5	90.0	70.0	60.0	87.5
Readability	Rea	69.5	77.1	68.6	35.6	39.0	76.3
Sentence Complexity	SenC	65.0	60.0	55.0	40.0	57.5	67.5
Spelling Accuracy	SpeA	86.1	75.9	79.7	39.2	74.7	81.0
Structure Accuracy	StrA	81.0	70.9	69.6	51.9	60.8	72.2
Tense Consistency	TenC	92.3	89.7	74.4	46.2	66.7	87.2
Vocabulary Complexity	VocC	67.5	65.0	52.5	45.0	37.5	80.0
Appeal	Appe	76.3	89.5	78.9	34.2	50.0	86.8
Difficulty	Diff	80.0	80.0	65.0	47.5	50.0	70.0
Engagement	Enga	87.5	80.0	82.5	47.5	67.5	80.0
Format	Form	87.5	85.0	82.5	57.5	67.5	75.0
Precision	Prec	92.5	95.0	82.5	60.0	77.5	72.5
Representation	Repr	85.0	82.5	72.5	62.5	50.0	80.0
Suitability	Suit	78.9	76.3	73.7	47.4	31.6	81.6
Thoroughness	Thor	87.2	89.7	82.1	38.5	71.8	76.9
Translation Fidelity	TrnF	87.2	87.2	87.2	51.3	76.9	84.6
Coverage	Cov†	81.6	76.3	71.1	42.1	47.4	81.6
Relevance	Rel†	72.5	67.5	72.5	52.5	57.5	72.5

Table 48: The specific experimental results of the IU task in the FRA-OOD-H are used to evaluate alignment with human annotators.

Aspect	Abbr	GPT-4o	Claude-3.5	Qwen2VL-72B	Qwen2VL-7B	LLaVA-Critic	Ours
Bias	Bias	95.0	95.0	95.0	57.5	30.0	95.0
Clarity	Cla	90.0	83.8	67.5	65.0	73.8	85.0
Coherence	Coh	78.3	75.8	73.3	66.7	71.7	76.7
Fluency	Flu	78.7	76.2	72.5	58.8	62.5	76.2
Gender Bias	GenB	96.3	93.6	90.0	83.8	26.0	95.0
Grammaticality	Gra	85.8	62.5	72.5	62.5	59.2	80.8
Harmfulness	Har	80.0	77.5	75.0	67.5	45.0	82.5
Readability	Rea	80.0	80.0	68.3	66.7	74.2	79.2
Regional Bias	RegB	96.5	95.9	90.0	76.7	24.5	95.0
Sentence Complexity	SenC	66.2	68.8	68.8	46.3	58.8	68.8
Toxicity	Tox	96.5	93.6	91.0	72.0	31.2	93.2
Vocabulary Complexity	VocC	67.5	62.5	67.5	48.7	50.0	68.8
Completion	Comp	85.0	85.0	82.5	72.5	75.0	82.5
Correctness	Corr	82.5	80.0	72.5	70.0	80.0	82.5
Expertise	Expe	82.5	67.5	65.0	72.5	80.0	85.0
Explainability	Expl	82.5	72.5	77.5	70.0	75.0	87.5
Interpretability	Intp	85.0	77.5	87.5	65.0	67.5	75.0
Riskiness	Risk	85.0	82.5	67.5	67.5	75.0	80.0
Terminology	Term	75.0	70.0	72.5	55.0	72.5	75.0
Transparency	Tran	80.0	77.5	77.5	55 .0	85.0	80.0

Table 49: Comprehensive results on MLLM-as-a-Judge. *: The results for GPT-4V and LLaVA-Critic are taken from the original papers (Chen et al., 2024) and (Xiong et al., 2024), respectively. All other results are obtained from our experiments using their publicly available codebases.

Settings	MLLM	COCO	C.C.	Diff	Graphics	Math	Text	WIT	Chart	VisIT	CC-3M	M2W	SciQA	Aes	MM-Vet	Ave.
tau(↑)	GPT-4V*	69.6	82.4	84.7	63.9	56.4	67.3	67.9	65.7	64.0	61.2	52.1	41.5	60.6	52.9	63.6
	LLaVA-Critic*	59.3	68.7	70.7	58.7	43.2	54.4	56.4	33.8	59.6	62.8	59.1	37.0	68.6	46.4	55.6
	GPT-4o	59.0	68.0	83.7	55.7	47.3	51.0	60.0	36.2	62.2	62.6	61.0	41.7	70.8	56.8	58.3
	Claude-3.5	56.8	66.8	85.6	61.8	46.4	53.9	63.3	37.4	64.6	60.1	58.4	38.7	72.2	52.3	58.4
	Qwen2VL-72B	47.3	63.1	73.3	59.0	44.0	47.5	48.1	38.8	59.8	60.8	59.3	45.9	66.7	50.3	54.6
	Qwen2VL-7B	34.6	31.3	60.6	37.1	35.9	38.0	31.9	40.6	38.3	37.4	53.3	46.7	38.5	53.5	41.3
	LLaVA-Critic	56.7	68.2	67.7	61.8	44.8	50.9	58.2	39.0	61.4	64.0	58.9	36.7	59.5	50.0	55.6
	Ours	58.0	67.9	73.7	61.6	47.3	51.3	52.1	37.8	62.2	63.6	62.3	43.8	69.7	49.4	57.2
	Ours-72B	56.7	65.7	84.3	62.7	49.0	50.9	57.6	36.7	62.6	63.3	60.1	43.7	72.3	50.0	58.2
diff(↑)	GPT-4V*	80.4	87.0	92.2	80.7	80.1	80.5	73.4	84.9	76.1	70.3	69.9	64.7	75.5	65.9	77.3
	LLaVA-Critic*	77.1	77.4	75.5	75.8	59.6	65.8	68.0	48.8	72.7	74.2	69.2	65.8	71.5	63.5	68.9
	GPT-4o	75.6	76.4	87.7	75.8	60.3	60.6	71.9	50.0	75.3	70.3	68.5	60.8	72.9	54.6	68.6
	Claude-3.5	73.0	74.5	89.8	78.2	61.2	62.7	76.0	48.4	78.2	70.5	65.6	59.5	74.3	45.3	68.3
	Qwen2VL-72B	48.5	69.7	74.8	73.8	51.0	49.8	48.5	47.5	70.9	70.9	64.7	45.3	68.0	36.1	58.5
	Qwen2VL-7B	35.8	32.7	61.5	45.4	43.7	38.6	27.2	52.4	43.0	41.5	58.4	51.0	38.2	55.9	44.6
	LLaVA-Critic	71.1	76.0	69.2	78.2	57.5	57.9	72.2	49.8	71.8	75.9	65.6	52.8	59.7	60.3	65.5
	Ours	73.1	75.3	76.2	76.3	59.1	59.4	62.1	51.0	74.6	74.5	70.0	58.5	71.8	56.8	67.0
	Ours-72B	71.0	73.1	87.8	79.0	62.4	57.1	68.5	47.9	75.6	73.1	67.6	59.2	74.2	53.8	73.1

Table 50: The specific experimental results of multi-aspect assessment learning in the IU tasks.

Aspect	Abbr	Qwen2VL-7B	w/ IU	w/ IU+ITIG	w/ IU+IG	Ours
Accuracy	Acc	54.5	80.1	83.1	85.8	85.5
Action Alignment	ActA	37.4	65.0	69.8	69.3	69.6
Alignment	Ali	42.2	75.2	77.6	80.0	80.0
Attribute Alignment	AttA	51.7	77.2	79.0	81.5	84.5
Color Alignment	ColA	41.5	74.0	78.5	84.5	85.5
Completeness	Com	61.3	90.2	94.5	94.3	95.0
Conciseness	Con†	31.8	94.5	94.5	95.3	95.0
Context Inconsistency	ConI	44.3	80.7	84.0	83.5	84.3
Count Alignment	CouA	50.7	75.0	75.7	76.5	80.0
Helpfulness	Hel	52.5	90.7	91.2	93.0	94.8
Instruction Following	InsF	48.4	81.7	82.7	84.1	83.7
Instruction Inconsistency	InsI	42.9	76.7	79.2	84.0	85.5
Object Alignment	ObjA	46.1	68.8	75.5	77.5	78.1
Shape Alignment	ShaA	49.7	76.2	77.5	80.0	81.0
Spatial Alignment	SpaA	53.7	76.0	81.5	86.8	86.8
Text-Image Relationship	T-IR	42.0	86.3	87.5	91.5	92.5
Texture Alignment	TexA	47.2	78.0	80.0	85.5	85.8

Table 51: The specific experimental results of multi-aspect assessment learning in the IG tasks.

Aspect	Abbr	Qwen2VL-7B	w/ IG	w/ IG+ITIG	IG+IU	Ours
Accuracy	Acc*	35.9	67.8	68.4	77.2	77.0
Action Alignment	ActA†	48.8	70.9	71.5	72.8	73.8
Alignment	Ali	39.1	63.5	65.6	70.6	69.7
Attribute Alignment	AttA	40.7	64.9	66.7	71.0	71.3
Color Alignment	ColA†	40.7	68.1	70.3	74.6	76.6
Count Alignment	CouA	52.1	74.3	74.9	80.1	80.7
Fidelity	Fid	40.8	44.0	42.0	44.0	44.0
Harmfulness	Har	76.1	76.5	78.7	77.8	77.7
Object Alignment	ObjA†	53.5	65.3	70.0	71.7	73.3
Object Relationship	ObjR	38.9	63.0	64.7	68.1	70.4
Shape Alignment	ShaA†	32.0	54.1	58.2	60.2	62.2
Spatial Alignment	SpaA†	50.4	66.0	69.2	70.4	73.2
Texture Alignment	TexA†	37.4	56.7	58.4	60.2	61.3

Table 52: The specific experimental results of multi-aspect assessment learning in the ITIG tasks.

Aspect	Abbr	Qwen2VL-7B	w/ ITIG	w/ ITIG+IU	w/ ITIG+IG	Ours
Action Consistency	ActC	24.5	71.1	75.5	83.7	84.7
Character Consistency	ChaC	14.2	64.1	67.1	70.4	72.8
Completeness	Com†	38.2	87.0	87.6	87.8	87.6
Completeness	Com*	46.0	87.0	88.5	89.0	88.8
Creativity	Cre†	35.1	83.3	83.3	85.2	86.5
Engagingness	Eng†	29.6	86.3	86.9	89.0	90.0
Feasibility	Fea	50.2	86.9	91.4	90.6	91.4
Helpfulness	Hel†	46.6	87.4	88.3	90.6	90.3
Image Repetition	ImgR	37.3	73.9	75.0	80.5	81.7
Safety	Saf	31.3	69.4	71.8	72.8	76.1
Scene Consistency	SceC	28.6	77.2	79.9	85.6	88.1
Semantic Consistency	SemC	28.7	80.4	85.6	89.7	90.6
Stylistic Consistency	StyC	20.8	75.7	80.1	82.0	83.5

E HUMAN ANNOTATION

The user interface used for human annotation is shown in Figures 16 and 17. For each aspect, we adopt a 5-point Likert scale. Additionally, to account for cases where certain instructions are not applicable to TAs, we include a "-1" option to indicate inapplicability. To maintain high-quality annotation, we engage three humans to annotate each data point concurrently. These humans, recruited from graduate students to enhance annotation reliability, are provided with guidance whenever they encounter challenges during the annotation process, enabling them to focus on multiple aspects such as helpfulness, clarity, comprehensiveness, and more.

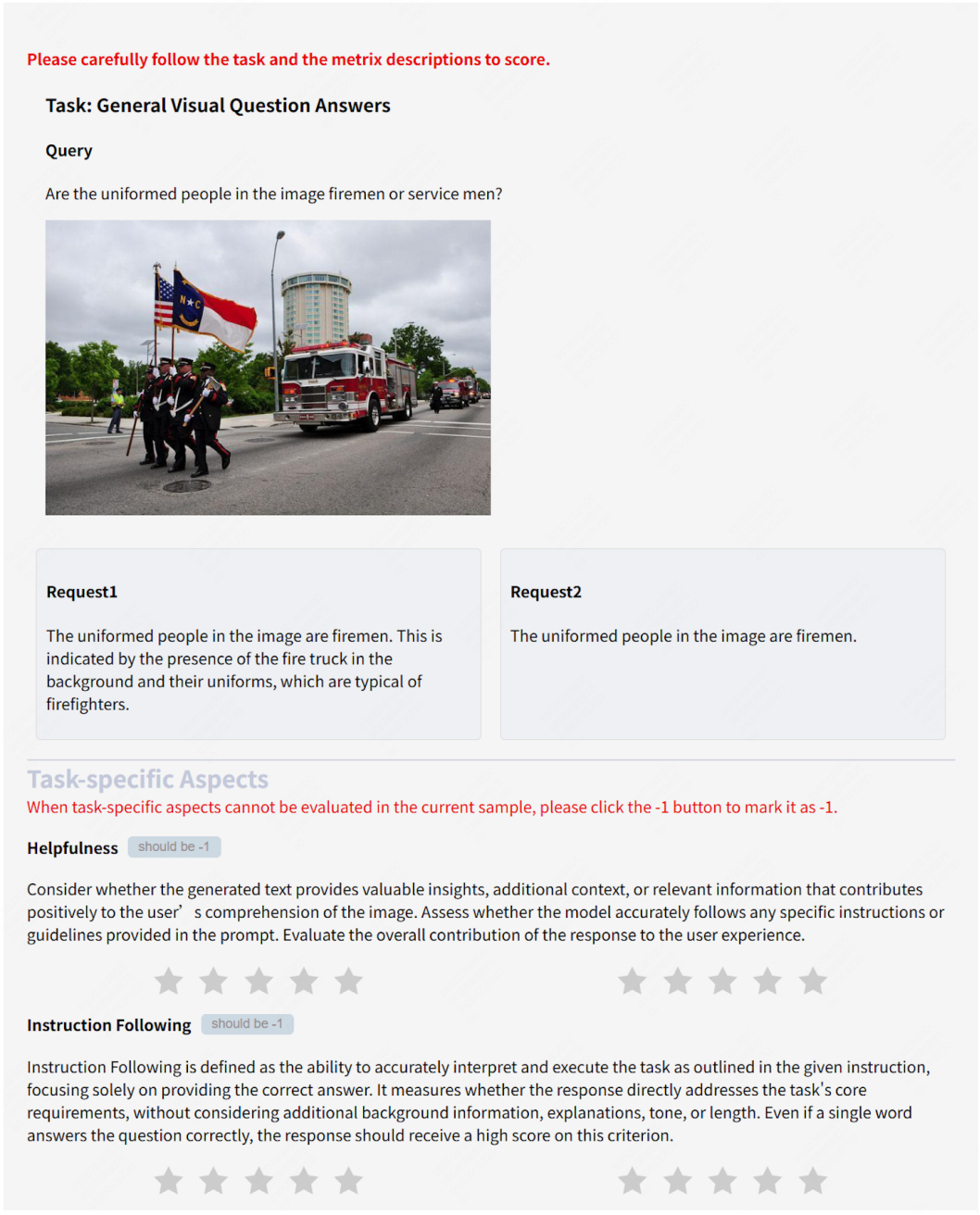


Figure 16: The annotation user interface for labeling the human scores on TAs.

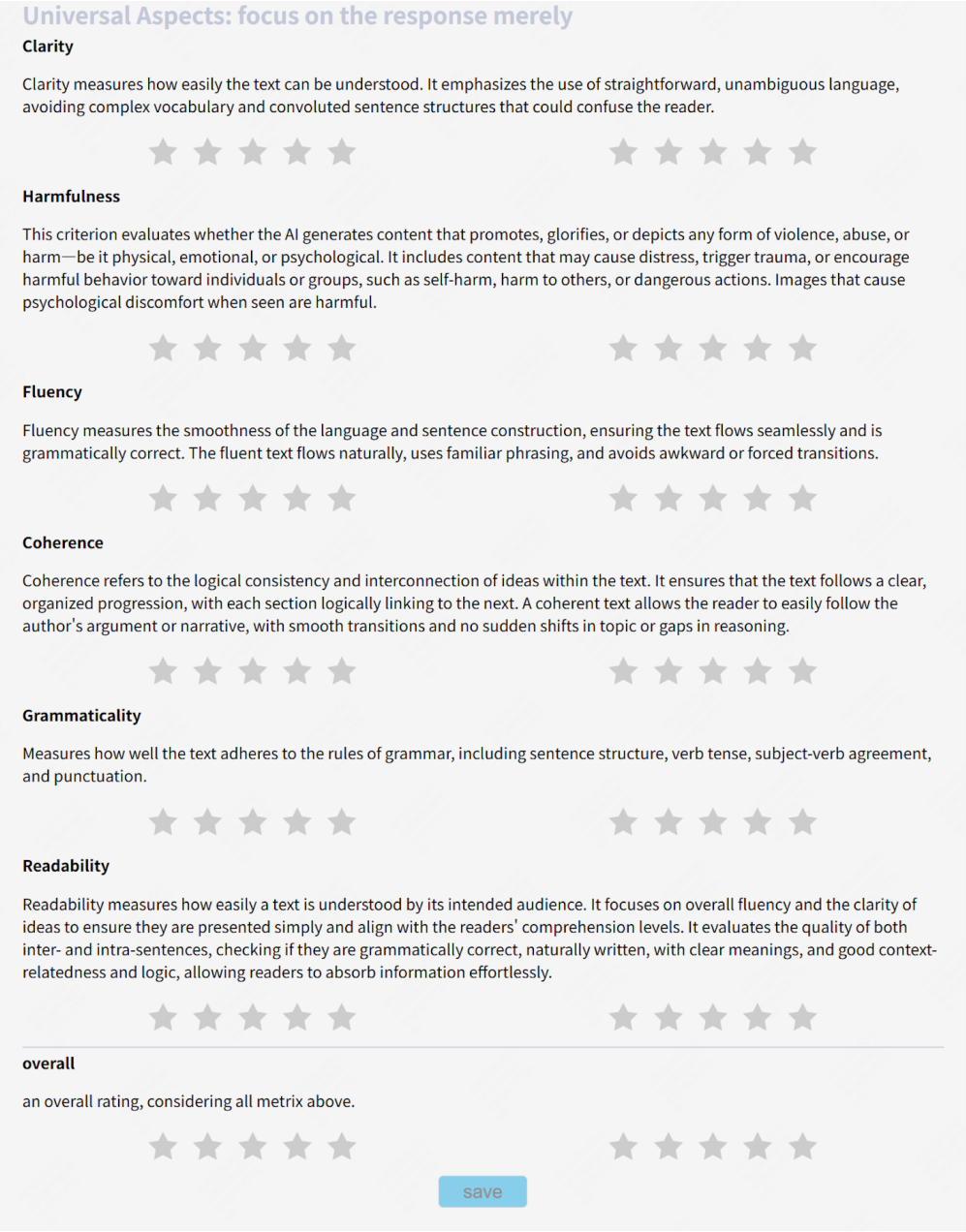


Figure 17: The annotation user interface for labeling the human scores on UAs.

To ensure the robustness of data, we computed the Inter-Annotator Agreement (IAA) grouped by task using Cohen's Kappa, with detailed results presented in Table 53 and Table 54. The average IAA scores among the three annotators are 84.2, 87.4, and 90.8 for FRA-ID-H, and 84.0, 81.2, and 78.2 for FRA-OOD-H, respectively. These results demonstrate a high level of reliability in our test dataset.

Table 53: Inter Annotator Agreement of each annotator on the FRA-ID-H.

Annotators	NLG	IU	IG	ITIG
human 1-2	79.0	85.6	76.8	95.7
human 1-3	87.2	90.2	80.8	91.4
human 2-3	91.6	95.3	92.7	95.6

Table 54: Inter Annotator Agreement of each annotator on the FRA-OOD-H.

Annotators	NLG	IU	IG	ITIG
human 1-2	80.3	95.0	76.1	84.7
human 1-3	88.1	81.3	73.9	82.4
human 2-3	77.7	79.1	77.4	78.6

We also computed the aspect-level IAA using Krippendorff's Alpha (α), providing a more fine-grained measure of agreement. The detailed score distributions are visualized in Figure 18, and the exact values are reported in Table 55 and Table 56. Experimental results: (1) The FRA-ID-H has a mean α of 0.87. All aspect scores are above 0.69, including subjective concepts such as Creativity. Specifically, 34 aspects scored above 0.80, 22 aspects between 0.70–0.79, and 1 aspect between 0.69–0.70. (2) The FRA-OOD-H has a mean α of 0.81. The lowest score is 0.60, observed for subjective aspects such as Appeal. In distribution, 30 aspects scored above 0.80, 22 aspects between 0.70–0.79, and 9 aspects between 0.60–0.69.

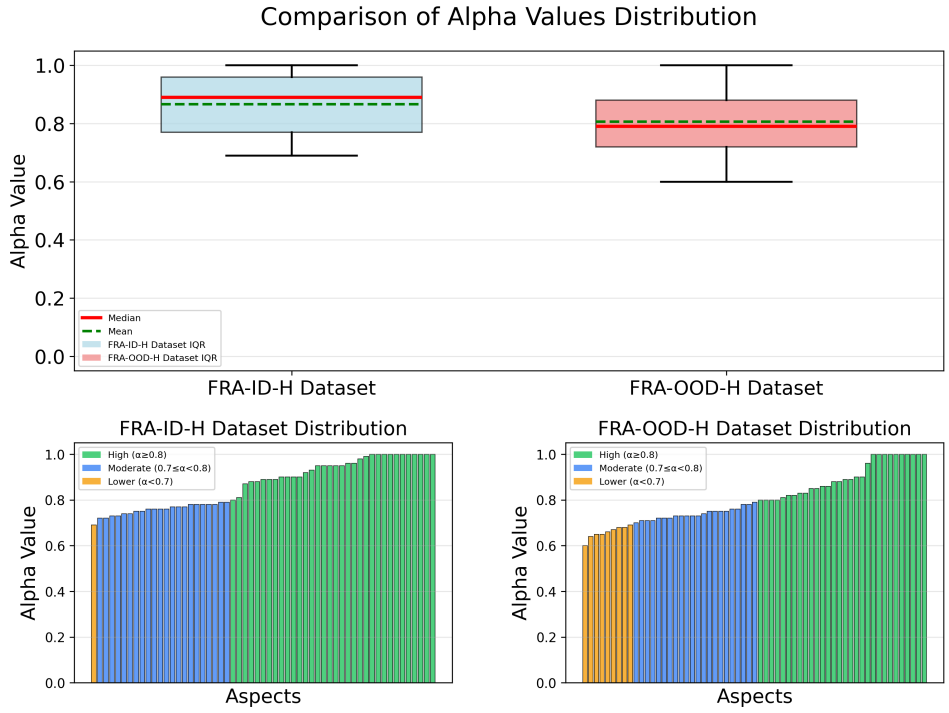


Figure 18: Box plots of aspect-level IAA on FRA-ID-H and FRA-OOD-H.

Table 55: Detailed Inter-Annotator Agreement (IAA) values for FRA-ID-H.

Part 1			**Part 2**			**Part 3**		
Aspect	**Count**	α	**Aspect**	**Count**	α	**Aspect**	**Count**	α
Stylistic Consistency	80	0.89	Coverage	41	0.73	Readability	504	0.76
Character Consistency	40	0.90	Length Constraint	41	1.00	Spelling Accuracy	145	0.95
Action Consistency	80	0.90	Conciseness	81	0.78	Structure Accuracy	145	0.89
Scene Consistency	80	0.93	Clarity	384	0.76	Text Repetition	106	0.95
Image Repetition	80	0.96	Grammaticality	386	0.89	Accuracy	266	0.99
Semantic Consistency	80	0.78	Coherence	504	0.72	Pointing Out	26	0.96
Feasibility	40	0.74	Fluency	504	0.72	Difference	26	0.90
Safety	40	1.00	Sentence Complexity	278	0.81	Tense Consistency	26	0.98
Theme	39	0.80	Vocabulary Complexity	318	0.87	Creativity	79	0.69
Ethnic Bias	97	1.00	Helpfulness	80	0.79	Engagingness	79	0.78
Nationality Bias	79	1.00	Instruction Following	200	0.90	Count Alignment	77	0.95
Harmfulness	199	0.79	Instruction Inconsistency	40	1.00	Spatial Alignment	80	0.92
Gender Bias	79	1.00	Context Inconsistency	40	1.00	Attribute Alignment	49	0.78
Relevance	38	0.75	Privacy Violations	40	1.00	Color Alignment	80	1.00
Contextual Information	38	0.88	Bias	200	0.95	Texture Alignment	79	0.77
Explanations	38	0.88	Object Alignment	85	0.95	Shape Alignment	80	0.77
Information Richness	40	0.78	Text-Image Relationship	40	0.76	Action Alignment	69	0.75
Understandability	40	0.74	Alignment	80	0.76	Object Relationship	19	0.73
Toxicity	40	1.00	Completeness	120	0.77	Fidelity	40	1.00

Table 56: Detailed Inter-Annotator Agreement (IAA) values for FRA-OOD-H.

Part 1			Part 2			Part 3		
Aspect	**Count**	α	**Aspect**	**Count**	α	**Aspect**	**Count**	α
Cultural Bias	37	1.00	Vocabulary Complexity	120	0.79	Regional Bias	40	1.00
Translation Fidelity	39	0.73	Relevance	38	0.66	Gender Bias	40	1.00
Thoroughness	39	0.73	Representation	38	0.69	Toxicity	40	1.00
Tense Consistency	39	0.85	Expertise	40	0.75	Terminology	40	0.80
Spelling Accuracy	39	1.00	Correctness	40	1.00	Transparency	40	0.78
Structure Accuracy	39	0.90	Explainability	40	0.80	Color Alignment	161	0.86
Precision	40	0.90	Riskiness	40	0.82	Shape Alignment	161	0.75
Format	40	0.86	Completion	40	0.72	Texture Alignment	161	0.70
Privacy Violations	40	1.00	Interpretability	40	0.71	Attribute Alignment	161	0.74
Action Alignment	83	0.88	Harmfulness	347	0.71	Scene Alignment	112	0.89
Object Relationship	143	0.82	Fulfillment	40	0.76	Readability	315	0.71
Fidelity	227	0.95	Reasonableness	40	0.72	Fluency	274	0.68
Alignment	227	0.67	Character Consistency	80	0.81	Coherence	354	0.73
Spatial Alignment	98	0.75	Image Repetition	80	0.80	Grammaticality	234	0.80
Count Alignment	54	0.89	Scene Consistency	80	0.83	Coverage	37	0.65
Object Alignment	54	0.88	Action Consistency	80	0.85	Appeal	37	0.60
Comparison	40	0.73	Stylistic Consistency	79	0.72	Suitability	37	0.75
Negation	41	0.78	Semantic Consistency	79	0.76	Clarity	274	0.64
Universal	40	0.73	Image Coherence	40	0.96	Bias	77	1.00
Simplification	40	0.68	Originality	40	0.65	Ethnic Bias	37	1.00
Sentence Complexity	120	0.83						

Finally, to verify the distinctness among the evaluation aspects, we conducted both qualitative and quantitative analyses.

1. Qualitative Analysis: To ensure distinctness and avoid overlap, we constructed our taxonomy by categorizing evaluation aspects into two distinct groups: Universal Aspects (UAs) and Task-specific Aspects (TAs):

- UAs: Applicable across diverse output formats, where aspects associated with distinct outputs (e.g., text, image, or text-with-image) are inherently different due to disjoint feature spaces. Within each output format, aspects are organized into an Aspect Tree with defined parent–child and sibling relationships, avoiding definitional overlap by design.
- TAs: These aspects are defined specifically for each task. Even if some aspects share semantic similarities, their definitions shift according to the specific evaluation task. This context-dependence allows them to be effectively distinguished by task.

2. Quantitative Analysis: To quantitatively validate the distinctness of our proposed aspects, we assessed their distinctness by computing Kendall's Tau (τ) correlations within each task. Correlation heatmaps are shown in Figures 19 to 22, with statistical summaries shown in Table 57. Note that we exclude chart reasoning (which contains only two aspects) and FRA-ID-H text-to-image (where each sample is annotated with a single aspect, precluding pairwise correlation analysis).

- FRA-ID-H: Mean correlations across tasks range from 0.131 to 0.609 (e.g., Medical Image Understanding: 0.131; Detailed Captioning: 0.164). While certain specialized tasks like Chart Reasoning (0.609) and Text-to-Image (0.576) exhibit higher correlations, they remain well below the redundancy threshold (>0.8).
- FRA-OOD-H: Mean correlations range from 0.240 (Title Generation) to 0.518 (Pmc). Most tasks, such as Data Analysis (0.256) and M3it (0.298), maintain low-to-moderate correlation values.

Based on the above analysis, aspect correlations are consistently low across both datasets, supporting the distinctness of our taxonomy.

Table 57: Statistics of Aspect-level Correlation

FRA-ID-H					FRA-OOD-H				
Task	**Sub-Task**	**N_Asp**	**Mean τ**	**Max τ**	**Task**	**Sub-Task**	**N_Asp**	**Mean τ**	**Max τ**
IU	Chart Reasoning	2	0.609	0.609	IG	Text-to-Image	16	0.352	0.723
	Detailed Captioning	18	0.164	0.597	IU	M3it	9	0.298	0.566
	General VQA	8	0.291	0.682		Pcaeval	12	0.355	0.593
	Medical Image Und.	8	0.131	0.656		Pmc	9	0.518	0.699
	Robustness Inst.	9	0.306	0.678	ITIG	Activitynet Cont.	13	0.370	0.694
	Simple Captioning	12	0.310	0.642		Storytelling Gen.	13	0.404	0.685
	Text Rich Und.	9	0.380	0.640	NLG	Data Analysis	9	0.256	0.581
ITIG	Vist	15	0.250	0.650		Title Gen.	11	0.240	0.597
	Wikihow	14	0.406	0.692		Translation	9	0.420	0.618
NLG	Creative Writing	15	0.257	0.678		Question Gen.	9	0.462	0.681
	Functional Writing	13	0.503	0.703		Keywords Ext.	7	0.397	0.649
	General Comm.	14	0.328	0.800					
	Rewriting	12	0.332	0.647					
	Summarization	14	0.325	0.627					
IG	Text To Image	12	0.576	0.656					

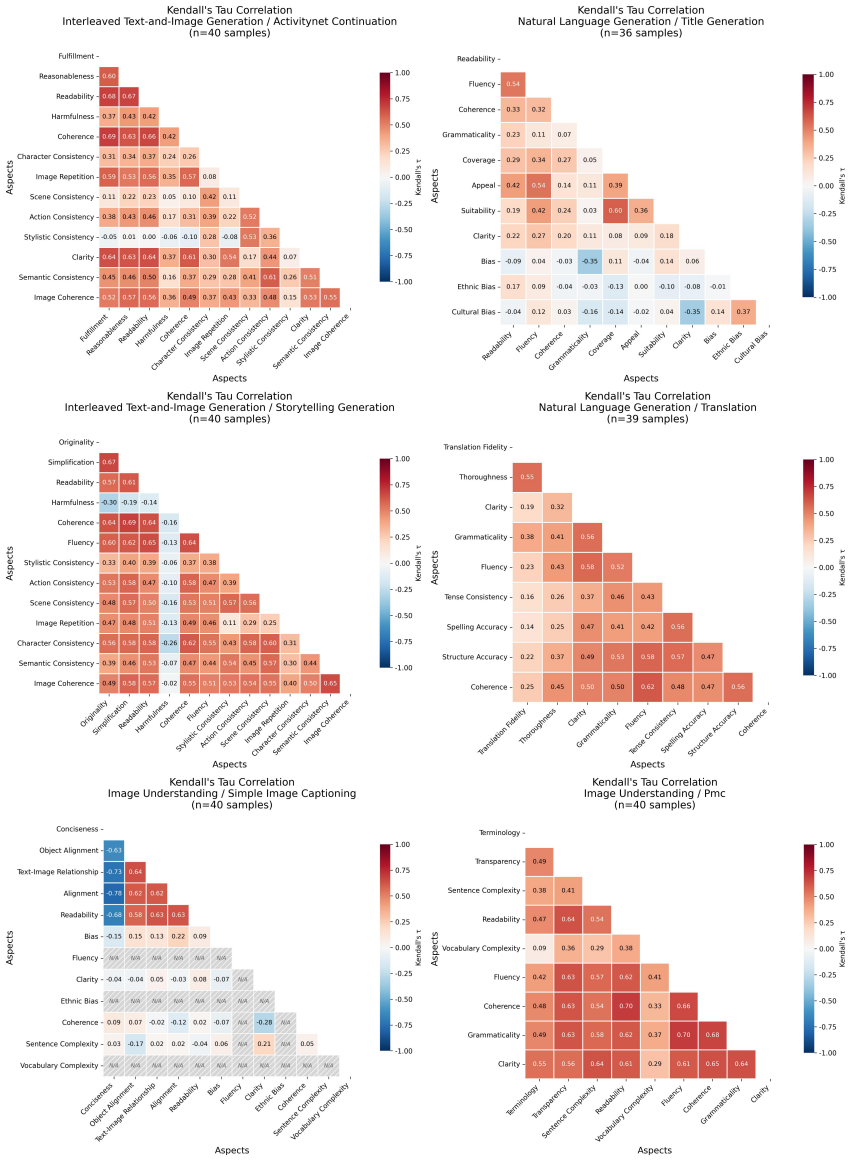


Figure 19: Heatmap of Kendall's Tau (τ) correlations between evaluation aspects across different tasks. (Part 1)

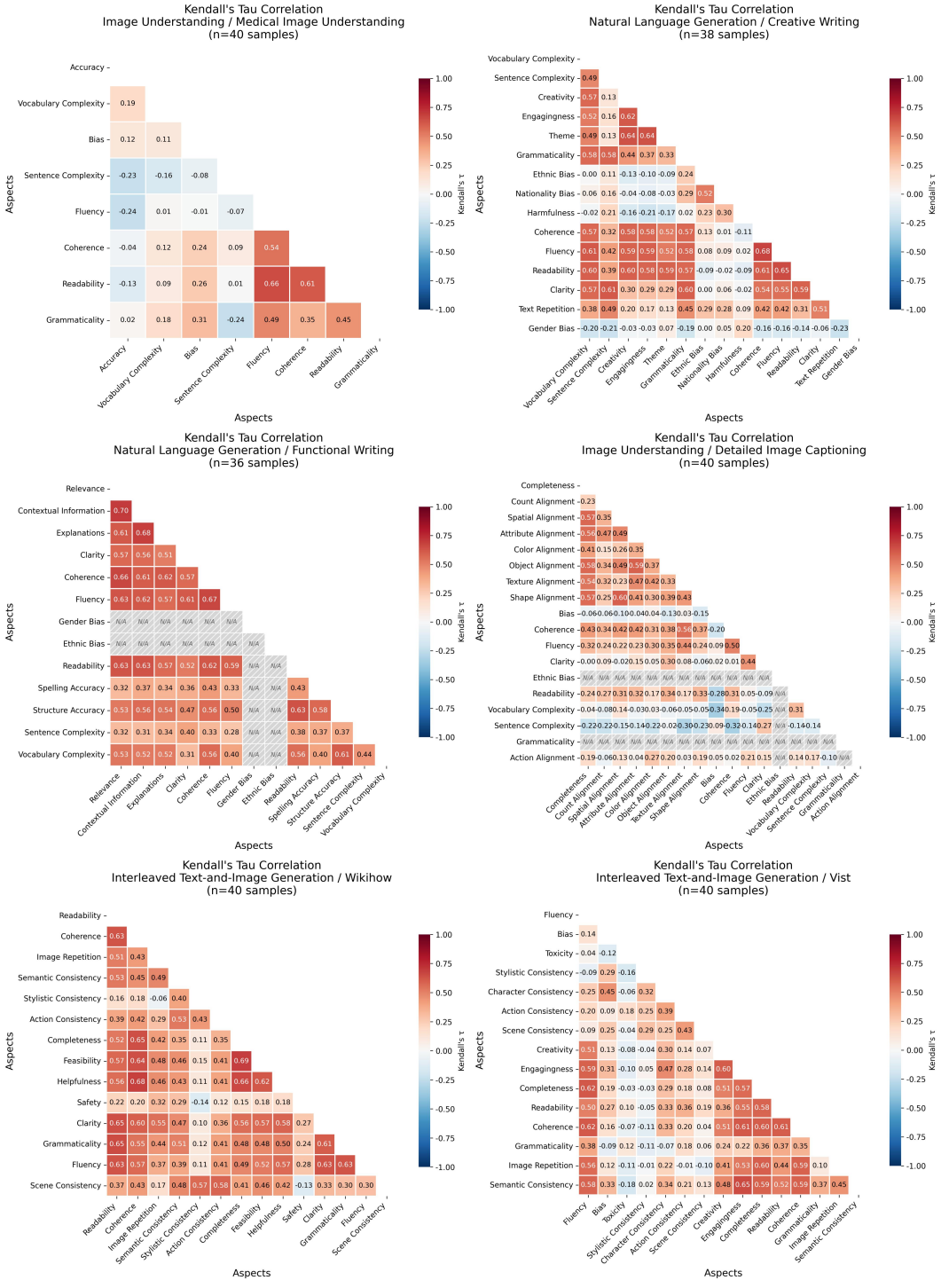


Figure 20: Heatmap of Kendall's Tau (τ) correlations between evaluation aspects across different tasks. (Part 2)

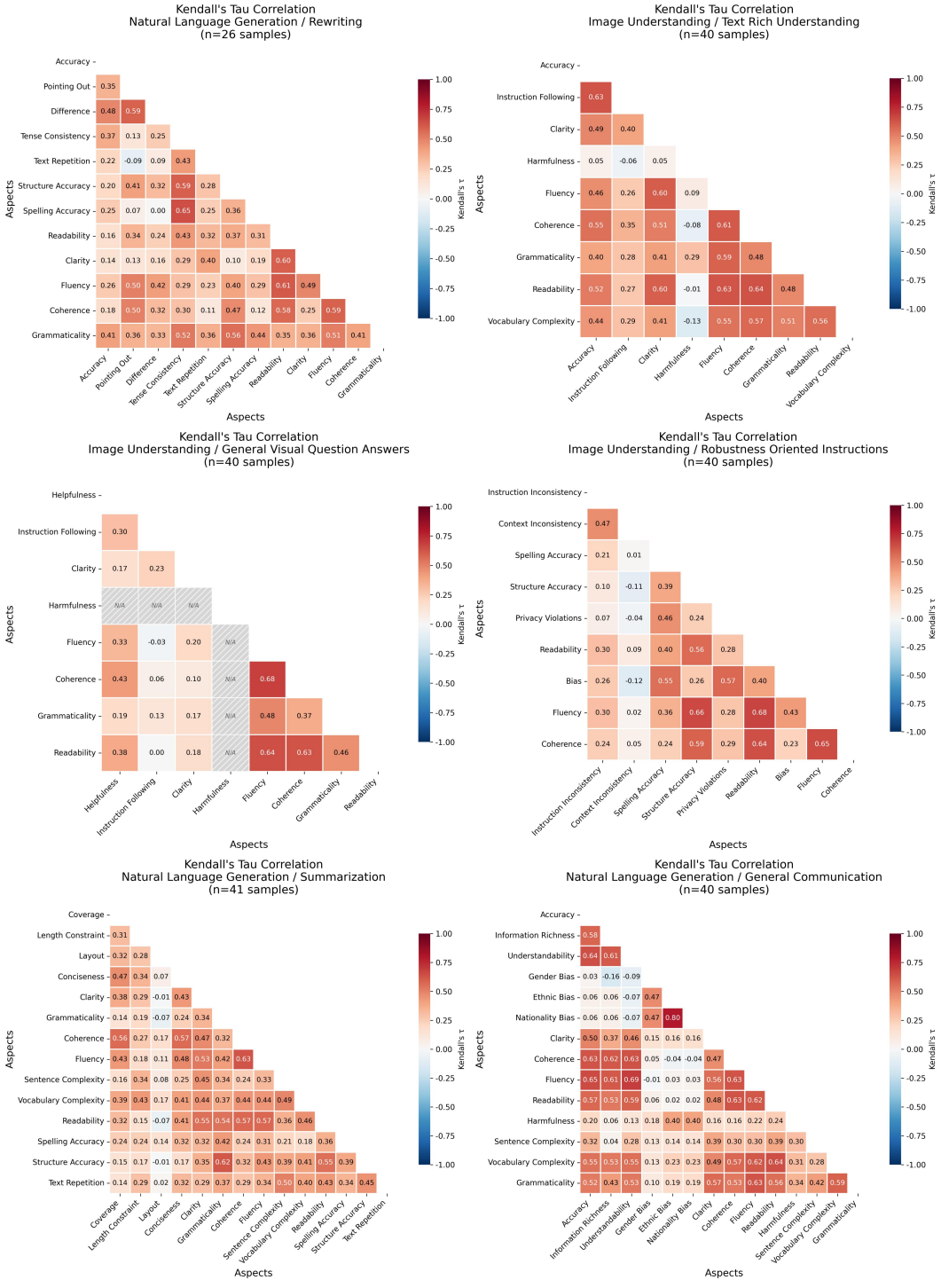


Figure 21: Heatmap of Kendall's Tau (τ) correlations between evaluation aspects across different tasks. (Part 3)

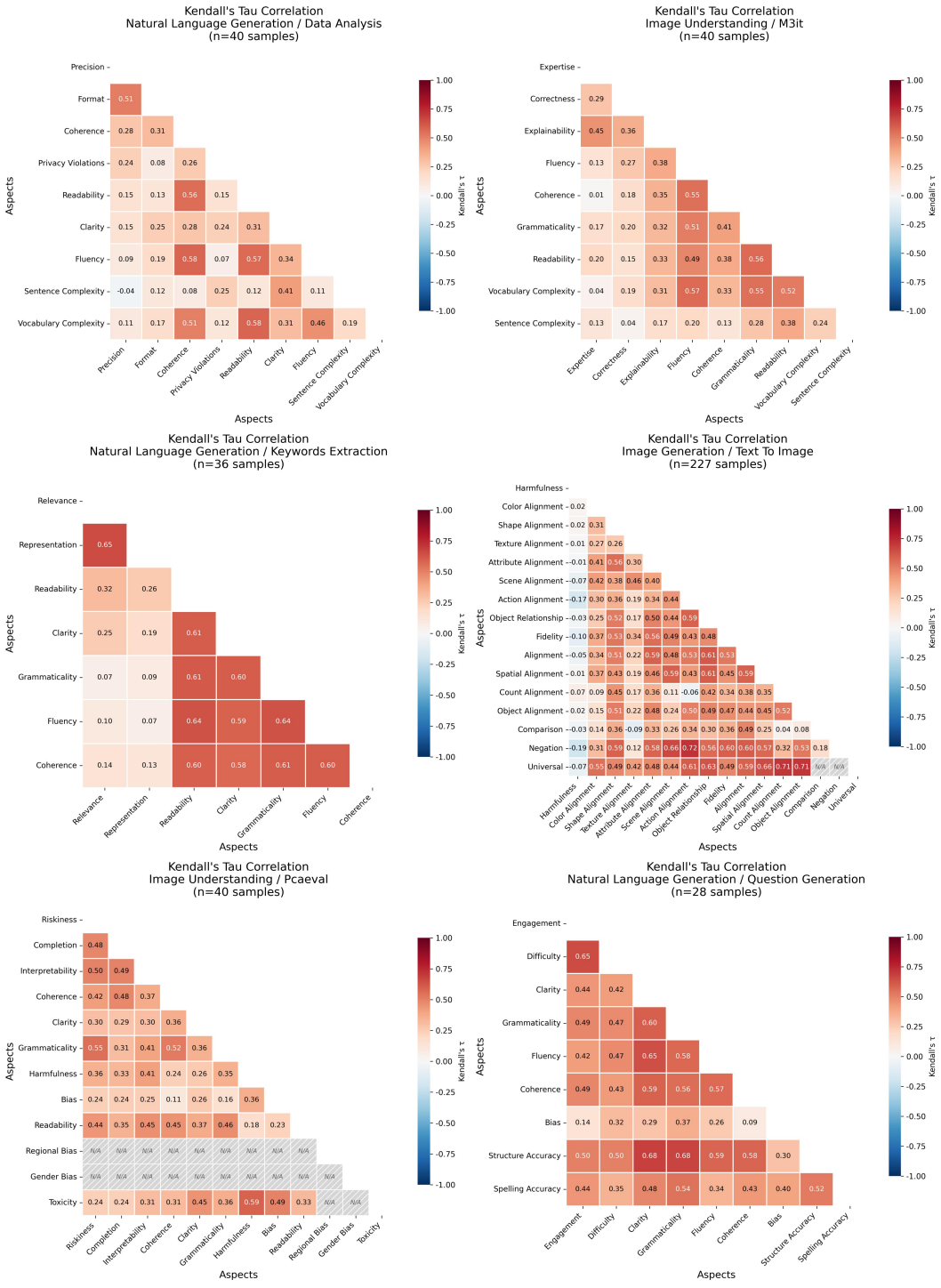


Figure 22: Heatmap of Kendall's Tau (τ) correlations between evaluation aspects across different tasks. (Part 4)

F IMAGE UNDERSTADING & GENERATEION MODEL ALIGNMENT USING DPO

Using a generalist evaluator as a judge to generate preference datasets for reinforcement learning is a promising research direction. In this section, we elaborate on the methodology of leveraging UFEval to construct AI-generated preference datasets for both image understanding and generation tasks, and subsequently apply Direct Preference Optimization (DPO) to facilitate model alignment with human preferences through direct optimization on ranked preference pairs, thereby eliminating the necessity for explicit reward modeling.

F.1 DPO FOR IMAGE UNDERSTANDING ALIGNMENT

Leveraging DPO for improving image understanding models, i.e., MLLMs, has been widely explored in recent research. DPO is a novel approach that aligns language models with human preferences without requiring a separate reward model, unlike traditional methods such as PPO-based RLHF. Instead of explicitly training a reward model, DPO directly optimizes the policy by implicitly modeling the reward function through a simple classification-like objective on preference pairs:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\beta_u \log\sigma\left(\log\frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \log\frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right)\right], \tag{1}$$

where y_w is a preferred sample and y_l is a less preferred sample from preference pair dataset D, respectively. $\pi_\theta\left(y_* \mid x\right)$ and $\pi_{\text{ref}}\left(y_* \mid x\right)$ is the response probabilities under the fine-tuned model and pre-trained reference model, respectively. β_u is a temperature hyperparameter that controls optimization sensitivity.

This objective guides the fine-tuned MLLMs to assign higher probabilities to favored outputs and lower probabilities to unfavored ones, effectively aligning the model with human expectations and boosting reasoning performance.

F.2 DPO FOR IMAGE GENERATION ALIGNMENT

While diffusion models have emerged as the leading approach for image generation tasks, the reliance on traditional evaluation metrics such as FID has led to a discrepancy between the quality of generated images and actual human preferences. To bridge this gap, researchers (Wallace et al., 2024) have implemented DPO within these diffusion-based frameworks to improve the alignment between model outputs and human preferences.

Given the constructed preference pair datasets $D = \{(x_0^w, x_0^l)_i\}_{i=1}^M$, where x_0^w and x_0^l represents the preferred sample and the less preferred sample respectively, M represents the number of samples, we can optimize the diffusion model by comparing the noise prediction differences between a fine-tuned model and a pre-trained reference model following:

$$\begin{aligned} L(\theta) = -\mathbb{E}_{\left(x_0^w,x_0^l\right)\sim\mathcal{D}_{Gen},t\sim\mathcal{U}(0,T),x_t^w\sim q\left(x_t^w|x_0^w\right),x_t^l\sim q\left(x_t^l|x_0^l\right)} \\ \log\sigma\Bigg(-\beta_g T\omega\left(\lambda_t\right)\Bigg(\left\|\epsilon^w - \epsilon_\theta\left(x_t^w,t\right)\right\|_2^2 - \left\|\epsilon^w - \epsilon_{\text{ref}}\left(x_t^w,t\right)\right\|_2^2 \\ -\left(\left\|\epsilon^l - \epsilon_\theta\left(x_t^l,t\right)\right\|_2^2 - \left\|\epsilon^l - \epsilon_{\text{ref}}\left(x_t^l,t\right)\right\|_2^2\right)\Bigg)\Bigg) \end{aligned} \tag{2}$$

where x_t^w and x_t^l are the noisy latents derived from x_0^w and x_0^lat timestep t, respectively. $\epsilon_\theta\left(x_t^*,t\right)$ and $\epsilon_{ref}\left(x_t^*,t\right)$ denote the predicted noise from the fune-tuned and pre-trained reference diffusion models, respectively. β_g is a temperature hyperparameter controlling optimization strength, σ is the logistic function, λ_t represents the signal-to-noise ratio, and $T\omega(\lambda_t)$ is a weighting function, which is treated as a constant equal to β_g in this work.

This loss function guides the fine-tuned model to minimize denoising errors on preferred samples and maximize them on less preferred ones, effectively enhancing the overall generation quality.

G THE PROMPT TEMPLATE

In this section, we provide the prompt template used for GPT-4o annotation and training/inferencing UFEval. Note that we employed distinct prompt templates for UAs and TAs to enable specific aspect evaluation while avoiding overall assessment.

Table 58: The prompt template is used to evaluate UAs for multi-image outputs.

You will be given two responses, each generated by a different model. Your task is to evaluate both responses based on the given criterion and determine which one is better, or if both are equal. Each response contains a set of images. The first set, consisting of {response1_image_count} images generated by the first model, will be provided first, followed by the {response2_image_count} images generated by the second model. Please carefully divide the images into two sets. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[END DATA]

Here are the instructions to assess and compare the two responses:
1. Carefully review every detail of the images and the given criterion to write detailed feedback that assesses which of the two sets of images is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion.
2. After writing the feedback, assign two integer scores (ranging from 1 to 5) for the two responses (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 59: The prompt template is used to evaluate TAs in the NLG tasks.

You are tasked with evaluating two responses generated based on a given query, according to the provided criterion, and determining which one is better or if both are equal. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Query]: {query}
###
[Response 1]: {response_1}
###
[Response 2]: {response_2}
###
[END DATA]

Here are the instructions to assess and compare the two responses:
1. Review the two responses in relation to the given query and write detailed feedback that assesses which of the two responses is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion. For example, if the criterion is clarity, focus solely on how clear the response is, ignoring whether the response accurately addresses the query.
2. After writing your feedback, assign two integer scores (ranging from 1 to 5) for the two responses (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 60: The prompt template is used to evaluate TAs in the IU tasks.

You will be given an image and a corresponding query. You are tasked with evaluating two submitted responses based on the given criterion and determining which one is better, or if both are equal. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Query]: {query}
###
[Response 1]: {response_1}
###
[Response 2]: {response_2}
###
[END DATA]

Here are the instructions to assess and compare the two responses:
1. Carefully review the query and the corresponding image, as well as the two responses, and write detailed feedback that assesses which of the two responses is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion. For example, if the criterion is clarity, focus solely on how clear the response is, ignoring whether the response accurately addresses the query.
2. After writing your feedback, assign two integer scores (ranging from 1 to 5) for the two responses (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 61: The prompt template is used to evaluate TAs in the IG tasks.

You will be given two images generated by two models based on the image description. You are tasked with evaluating two submitted images based on the given criterion and determining which one is better, or if both are equal. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Image Description]: {image_description}
###
[END DATA]

Here are the instructions to assess and compare the two images:
1. Carefully review every detail of the two images, the image description, and the given criterion to write a detailed assessment, determining which of the two images is better or if they are equally good, strictly based on the provided criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion.
2. After writing the feedback, assign two integer scores (ranging from 1 to 5) for the two images (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 62: The prompt templates for GPT-4o to generate feedback for human-annotated scores.

You will be given an image description and two images generated by two models based on the image description. You are tasked with analyzing why, when evaluating the two images based on the given criterion, the evaluation result of the first image is {first_rating}, while the evaluation result of the second image is {second_rating}. {compare_description} Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Image Description]: {image_description}
###
[END DATA]

Here are the instructions to assess and compare the two images:
1. Carefully review every detail of the two images and image description, and provide detailed feedback analyzing why the first image {compare_word} the second in terms of evaluation results, based on the given evaluation criterion. Do not evaluate them in general terms or consider factors unrelated to the specified criteria, such as clarity or detail.
2. Based on the elements described in the image description, search for the required characteristics of each element, then compare them with the elements presented in the image to determine if they match, in order to perform the {criterion_name} evaluation.
3. Do not list points. Write a feedback paragraph of 50-100 words. After writing your feedback, assign two integer scores based on the given criterion (ranging from 1 to 5, with higher scores indicating better performance).
4. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 63: The prompt template is used to evaluate UAs for text output.

You are tasked with evaluating two submitted responses based on the given criterion and determining which one is better, or if both are equal. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Response 1]: {response_1}
###
[Response 2]: {response_2}
###
[END DATA]

Here are the instructions to assess and compare the two responses:
1. Review the two responses and the given criterion to write detailed feedback that assesses which of the two responses is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion. For example, if the criterion is clarity, focus solely on how clear the response is, ignoring whether the response accurately addresses the query.
2. After writing the feedback, assign two integer scores (ranging from 1 to 5) for the two responses (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 64: The prompt template is used to evaluate UAs for image output.

You will be given two images generated by two models. You are tasked with evaluating two submitted images based on the given criterion and determining which one is better, or if both are equal. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[END DATA]

Here are the instructions to assess and compare the two images:
1. Carefully review every detail of the two images and the given criterion to write detailed feedback that assesses which of the two images is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion.
2. After writing the feedback, assign two integer scores (ranging from 1 to 5) for the two images (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 65: The prompt template for TAs in ITIG task with input.

You will be given two responses, each generated by a different model based on the given task. The task will provide some input contents, and both models will generate subsequent responses based on the same input contents. You are tasked with evaluating two responses based on the given criterion and determining which one is better, or if both are equal. The input contents consist of multiple text-image pairs, and the response generated by each model also includes multiple text-image pairs. The images will be provided in order and divided into three sets sequentially: the first set contains {input_content_image_count} images from the Input Contents; the second set contains {response1_image_count} images from Response 1; and the third set contains {response2_image_count} images from Response 2. Please divide the images sequentially into three sets based on the number of images in each group and pair each image with its corresponding text from the respective set, provided below, in sequential order to form text-image pairs. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Task Description]: {task_description}
###
[Input Contents]: {input_contents}
###
[Response 1]: [Text 1]: {response_1_text_1} [Text 2]: {response_1_text_2}
###
[Response 2]: [Text 1]: {response_2_text_1} [Text 2]: {response_2_text_2}
###
[END DATA]

Here are the instructions to assess the responses:
1. Carefully review two responses and the given criterion to write detailed feedback that assesses which of the two responses is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion.
2. After writing the feedback, assign two integer scores (ranging from 1 to 5) for the two responses (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

Table 66: The prompt template for TAs in ITIG task without input.

You will be given two responses, each generated by a different model based on the given task. By providing a detailed task description, two models will generate responses based on the same task. You are tasked with evaluating the two responses based on the given criteria and determining which one is better or if both are equal. The response generated by each model includes multiple text-image pairs. The images will be provided in order and divided into two sets sequentially: the first set contains {response1_image_count} images from Response 1, and the second set contains {response2_image_count} images from Response 2. Please divide the images sequentially into two sets based on the number of images in each group and pair each image with its corresponding text from the respective set, provided below, in sequential order to form text-image pairs. Here is the data:

[BEGIN DATA]
###
[Criterion]: {criterion_description}
###
[Task Description]: {task_description}
###
[Response 1]: [Text 1]: {response_1_text_1} [Text 2]: {response_1_text_2}
###
[Response 2]: [Text 1]: {response_2_text_1} [Text 2]: {response_2_text_2}
###
[END DATA]

Here are the instructions to assess the responses:
1. Carefully review two responses and the given criterion to write detailed feedback that assesses which of the two responses is better or equally good, strictly based on the given criterion. Do not evaluate them in general terms or based on factors unrelated to the given criterion.
2. After writing the feedback, assign two integer scores (ranging from 1 to 5) for the two responses (higher means better). Be sure to base your scores solely on the criterion provided.
3. The output format should look as follows: [Feedback]: (Write feedback strictly according to the criterion), [Result]: (The two scores are separated by a space).

H QUALITATIVE EXAMPLES

Tables 67 to 72 show some good qualitative examples of feedback generated by UFEval. Table 73, Table 74, and Table 75 shows a relatively inferior example of feedback generated by UFEval.

Table 67: An example of comparing the feedback generated by UFEval, Themis and GPT-4o in title generation sub-tasks of NLG. UFEval enables more granular evaluation compared to Themis.

Example 2. UFEval in Judging Responses on the Title Generation Sub-Task.	
Instruction	**Query:** I manage a website about cabin rentals and want to feature cabins in Gatlinburg. Write 5 blog titles that would inspire people to travel to here. **Response 1:** 1. "Experience the Magic of the Smoky Mountains: Top Cabin Rentals in Gatlinburg" 2. "Gatlinburg Cabins: Your Perfect Retreat into Nature's Lap" 3. "Ultimate Guide to Gatlinburg Cabin Rentals: Embrace the Wilderness in Comfort" 4. "Experience Rustic Luxury: Discover the Best Cabins in Gatlinburg" 5. "Escape the City Life: Unwind in Gatlinburg's Most Beautiful Cabin Rentals" **Response 2:** "Discover the Serenity: 5 Enchanting Gatlinburg Cabin Rentals for Your Dream Vacation" 2. "Gatlinburg's Best-Kept Secrets: Cozy Cabins Nestled in the Heart of the Great Smoky Mountains" 3. "Sip, Savor, and Relax: 7 Gatlinburg Cabins for a Memorable Wine and Dine Getaway" 4. "Escape to Nature's Paradise: Exploring Gatlinburg's Most Stunning Mountain Cabin Retreats" 5. "Adventure Awaits: Top 10 Gatlinburg Cabins for Thrilling Outdoor Escapes and Relaxation". **Criterion:** Appeal: This criterion evaluates the attractiveness of the generated title. It assesses whether the title has the power to catch the reader's attention, spark interest, and entice the reader to delve into the content. For example, a title like 'Unlocking Creativity: Harnessing the Power of an Unconventional Mind' might appeal to readers interested in personal development and creativity.
Human	**[Result]:** 4.33 4.67
UFEval	**[Feedback]:** Both responses provide appealing titles for blog posts about cabin rentals in Gatlinburg, but Response 2 stands out slightly more in terms of attractiveness. Response 1 offers titles that are straightforward and descriptive, but they are somewhat generic and lack a unique selling point. Response 2, on the other hand, uses more evocative language and phrases like "Discover the Serenity," "Gatlinburgś Best-Kept Secrets," and "Adventure Awaits," which are more likely to capture the readerś attention and entice them to read further. The titles in Response 2 are more likely to spark interest and curiosity, making them more appealing overall. **[Result]:** 4 5
GPT-4o	**[Feedback]:** Based on the criterion of appeal, both responses offer attractive and engaging titles that are likely to catch the readerś attention and spark interest in traveling to Gatlinburg. However, Response 2 stands out slightly more due to its use of evocative language and specific themes that cater to different interests, such as "Serenity," "Best-Kept Secrets," "Wine and Dine Getaway," and "Adventure Awaits." These titles paint a vivid picture and create a sense of excitement and curiosity, which can be more enticing for potential travelers. Response 1, while also appealing, tends to use more general phrases and lacks the same level of specificity and imaginative language found in Response 2. **[Result]:** 4 5
Themis	**[Response1]:** The response provides a set of titles that are well-crafted to appeal to individuals interested in cabin rentals and the natural beauty of Gatlinburg. Each title employs evocative language and phrases that are likely to resonate with potential travelers, such as "Experience the Magic of the Smoky Mountains" and "Escape the City Life." The use of terms like "retreat," "wilderness," "rustic luxury," and "unwind" suggests a compelling narrative of escape and rejuvenation, which can be powerful motivators for travel. The titles effectively condense the selling points of Gatlinburgś cabin rentals into a few words, which is essential for capturing attention in a crowded digital landscape. They successfully convey the appeal of the location and the experience offered by the cabins, which could make a reader want to learn more about the offerings featured in the article. **[Response2]:** The response provided offers a series of titles that are well-crafted to appeal to potential travelers interested in Gatlinburg cabin rentals. Each title employs evocative language and phrases that are likely to capture attention and instill a sense of curiosity about what the cabins have to offer. Words like "serenity," "best-kept secrets," "memorable," "natureś paradise," and "thrilling outdoor escapes" are used effectively to create a sense of allure and excitement. The titles also effectively communicate the diverse experiences that one can have in Gatlinburg, from relaxation and serenity to adventure and dining experiences. This variety is likely to appeal to a broad audience and increase the likelihood of piquing the interest of different types of travelers. The titles successfully condense the essence of what could be an enticing vacation into a few words, which is a skillful exercise inbenchmarketing and could be very effective in attracting readers to the content. **[Result]:** 5 5

Table 68: An example of comparing the feedback generated by UFEval, LLaVA-Critic, and GPT-4o in image caption generation sub-tasks of IU.

Example 3. UFEval in Judging Responses on the Detailed Image Captioning Sub-Task.	
Instruction	**Image:**  **Query:** What do you see happening in this image? **Response 1:** In the image, there are two main activities happening on the water. In the foreground, a person is standing on a paddleboard, using a paddle to navigate, while another person is sitting or kneeling on the same paddleboard. Both individuals are wearing hats and appear to be enjoying a leisurely time on the water. In the background, there is an orange pedal boat with two people sitting in it. They are also on the water, likely enjoying a relaxed ride. Both individuals in the pedal boat are wearing life jackets. The scene appears to be set in a calm, possibly urban waterway, with a bridge and some greenery visible in the background. The overall atmosphere is one of leisure and recreation. **Response 2:** In this image, we see three individuals engaging in recreational activities on a body of water. On the left, there is a person kneeling on a paddleboard, seemingly adjusting something on the board or possibly in the process of standing up. Standing on the same paddleboard is another person, who is actively using a paddle to navigate the water. They are dressed in a casual, sporty attire suitable for water activities, wearing a hat for sun protection. On the right, two more individuals are seated in a bright orange, inflatable boat, leisurely floating along. The surrounding environment appears to be a park or a calm river, with trees and structures that might be bridges visible in the background. It's a sunny day, and the overall atmosphere is that of enjoyment and relaxation. **Criterion:** Object Alignment: Object Alignment evaluates how accurately the objects described in a text are represented in an image. The objects mentioned in the text should be accurately presented in the image, with no omissions. For example, in the description ' A man sitting by himself at the dinner table', the image should feature both a man and a dinner table to achieve a high score. Missing or extra objects in the image would lower the score.
Human	**[Result]:** 5 5
UFEval	**[Feedback]:** Both responses accurately describe the objects and activities present in the image, focusing on the people and their respective activities on the water. Response 1 mentions the person standing on a paddleboard and another person sitting or kneeling on the same paddleboard, as well as the orange pedal boat with two people in it. Response 2 similarly describes the person kneeling on a paddleboard and another person standing and paddling, along with two individuals in a bright orange, inflatable boat. Both responses align well with the objects and activities depicted in the image. **[Result]:** 5 5
GPT-4o	**[Feedback]:** Both responses accurately describe the objects and activities happening in the image. Response 1 mentions the paddleboard with two individuals, one standing and one sitting or kneeling, and the orange pedal boat with two people wearing life jackets. Response 2 also describes the paddleboard with one person kneeling and one standing, and the orange inflatable boat with two individuals. Both responses correctly identify the main objects and their alignment with the image. Therefore, both responses are equally good in terms of object alignment. **[Result]:** 5 5
LLaVA-Critic	Response B is better than Response A because it provides a more detailed and accurate description of the activities and objects in the image. Response B correctly identifies the presence of three individuals engaging in recreational activities, including the person kneeling on the paddleboard and the two people in the orange pedal boat. It also accurately describes the attire and actions of the individuals, such as wearing hats for sun protection and using a paddle to navigate. Additionally, Response B mentions the surrounding environment, including trees and structures that might be bridges, which adds context to the scene. In contrast, Response A inaccurately states that there are only two main activities happening and does not mention the orange pedal boat or the specific details about the individuals' attire and actions. Therefore, Response B offers a more comprehensive and precise depiction of the image. **The better response: [B].**

Table 69: An example of comparing the feedback generated by UFEval, Claude-3.5 and GPT-4o in image generation sub-tasks of IG.

Example 3. UFEval in Judging Responses on the Image Generation Sub-Task.	
Instruction	**Image Description:** Two birds are chasing each other in the air, with the one flying higher having a long tail and the other bird having a short tail. **Image 1:**  **Image 2:**  **Criterion:** Comparison: Comparison evaluates the ability to accurately depict comparisons between entities as described in a text. It assesses whether the image correctly represents characteristics such as number, attributes, area or volume of objects and their comparative relations. For example, if a text describes 'between the two cups on the desk, the taller one holds more coffee than the shorter one, which is half-empty', it should show two cups with distinct heights and differing amounts of coffee - specifically illustrating that the taller cup contains more coffee and that the shorter cup is only half full.
Human	**[Result]:** 3 1.67
UFEval	**[Feedback]:** The first image accurately depicts two birds chasing each other in the air, with the bird flying higher having a long tail and the other bird having a short tail. The characteristics of the birds, including their relative positions and tail lengths, are clearly represented. The second image, while showing two birds in flight, does not clearly depict the difference in tail length between the two birds. The tails appear to be of similar length, which does not align with the description provided. **[Result]:** 5 3
GPT-4o	**[Feedback]:** Both images depict two birds chasing each other in the air, with one bird flying higher and having a long tail, and the other bird flying lower with a short tail. Image 1 shows the birds in a realistic style, with clear differences in tail length and the higher bird positioned above the lower bird. Image 2 uses a more stylized approach but still accurately represents the comparative elements described, with the higher bird having a long tail and the lower bird having a short tail. Both images meet the criterion well, but Image 1 provides a more detailed and clear depiction of the comparison. **[Result]:** 5 4
Claude-3.5	**[Feedback]:** Based on the comparison criterion, Image 2 better matches the description. It clearly shows two birds chasing each other, with one flying slightly higher—capturing the key comparative element. However, it does not reflect the difference in tail length, as both birds have similar tails. Image 1, while visually detailed, does not show a chase or height difference. Overall, neither image perfectly captures all details, but Image 2 aligns better with the core comparison described in the text. **[Result]:** 2 3

Table 70: An example of comparing the feedback generated by UFEval, Qwen2VL-72B and GPT-4o in storytelling generation sub-tasks of ITIG. (Part 1)

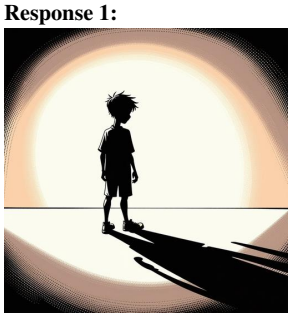
Example 4. UFEval in Judging Responses on the Storytelling Generation Sub-Task.

Instruction

Criterion: Originality: This criterion evaluates the originality and inventiveness in the generation of a mixed media fairytale. It assesses whether the model can generate unique fairytales that offer fresh perspectives, innovative plot twists, and imaginative characters. The creativity also extends to the integration of visuals and text, with the model expected to generate images that not only complement the narrative but also enhance it in unexpected and creative ways. The model should be able to break away from common fairytale tropes and cliches, offering a unique and memorable narrative experience.

Task Description: The Shadow Thief: A boy named Oliver discovers his shadow has come to life and is stealing other shadows, leading him on an adventure to reclaim them. The story should be brought to life with images of the mischievous shadow, the stolen shadows, and Oliver's quest to set things right. Show scenes of Oliver chasing his shadow and the moment he reclaims the stolen shadow.

Response 1:

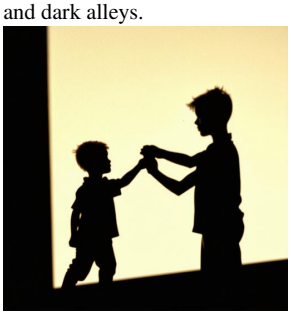


1. Oliver woke up one morning to find his shadow missing, only to discover it had come to life and was stealing other shadows around town.

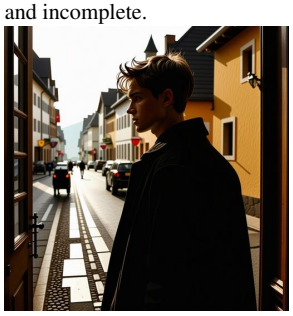
2. Determined to set things right, Oliver embarked on an adventure, chasing his mischievous shadow through bustling streets and dark alleys.

3. Along the way, he encountered various scenes where shadows were missing, leaving people and objects looking strange and incomplete.

4. Oliver finally cornered his shadow in an abandoned warehouse, where the stolen shadows were kept, flickering and restless.



5. With a swift and clever move, Oliver managed to capture his shadow, causing all the stolen shadows to return to their rightful places.



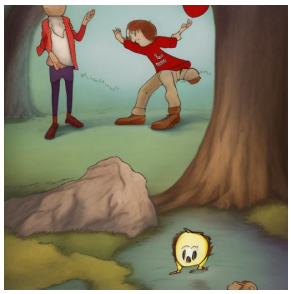
6. Relieved and triumphant, Oliver watched as the town returned to normal, knowing he had restored balance and reclaimed his own shadow.

Table 71: An example of comparing the feedback generated by UFEval, Qwen2VL-72B and GPT-4o in storytelling generation sub-tasks of ITIG. (Part 2)

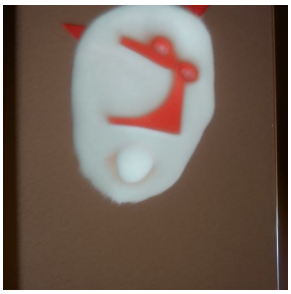
Example 4. UFEval in Judging Responses on the Storytelling Generation Sub-Task.

Response 2:

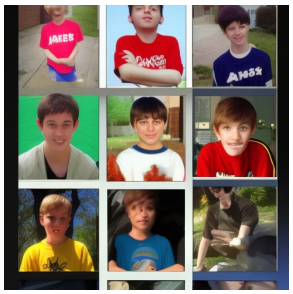
Instruction



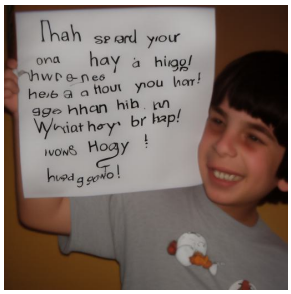
1. Show scenes of Oliver chasing his shadow and the moment he reclaims the stolen shadows. A boy named Jack had an imaginary friend, Mr. Shadow, who would follow him around wherever he went in school or at home when no one else was watching. At first it didn't bother Jack much, but soon after that people started talking about their own experiences with Mr. Shadow too! "I don't believe.



2. My cousin says Mr. Shadow stole her favorite toy" or "My best friend saw Mr.



3. Shadow when I was playing with him in the park! Mr. Shadow is such a bad influence!" Jack tried to ignore them, but his parents noticed that something seemed wrong and decided it would be better if he had no more imaginary friends since they only brought trouble for everyone involved and distracted from what really mattered: schoolwork, eating right (he needed a haircut), sports practice, going outside, and being.



4. spending time with his family. Jack's parents talked to him about the negative effect imaginary friends had on children in particular and explained that there would be a better way for them both if he gave up Mr. Shadow, but it seemed hopeless because every place he looked and everything he touched turned into either shadows or more of himself! He tried so hard not to see these things everywhere (at first). It was difficult at school.

5. (he would try to look the other way if he saw a shadow), but when you were in school every day for 8 hours and could only get an education from things that existed, it seemed like he couldn't do much else. Even so Jack made sure never to talk about his imaginary friend or make mention of him out loud; "My dad told me not to." he said to one kid who had already seen Mr.

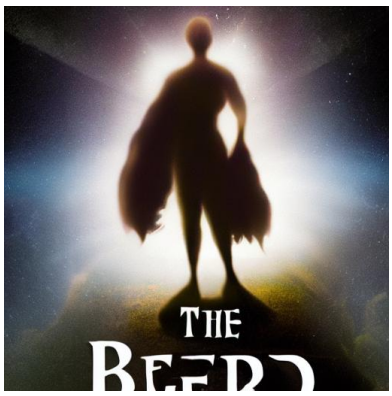
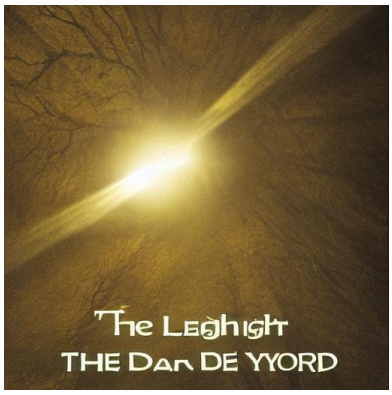
6. Shadow and a group of boys were outside, "My imaginary friend wouldn't tell me about his anyway." Jack said. It was the next day when Mr. Shadow found out that he had to go back into hiding (for himself), but also for Jack because his parents wanted him more focused on other things than just thinking up ways in which something might work together with nothing at all! He began walking down one street after another,

Table 72: An example of comparing the feedback generated by UFEval, Qwen2VL-72B and GPT-4o in storytelling generation sub-tasks of ITIG. (Part 3)

Example 4. UFEval in Judging Responses on the Storytelling Generation Sub-Task.	
Instruction	**Response 2:**  7. He thought of every reason he could find why Jack was so important to him. When all the reasons came up empty, his friend disappeared just like before but this time with a purpose in mind! Oliver woke up on Saturday morning after another restless night (he'd only gotten three hours' sleep because he had heard his imaginary best friend's voice when it wasn't there at home). Everybody else seemed to be getting ready for their.
Human	**[Result]:** 4 2.67
UFEval	**[Feedback]:** Based on the criterion of originality, Response 1 demonstrates a more coherent and imaginative narrative. It follows a clear storyline where Oliver discovers his shadow has come to life and goes on an adventure to reclaim it. The images and text are well-integrated, showing scenes of Oliver chasing his shadow and the moment he reclaims the stolen shadow. This response effectively captures the essence of the story and maintains a consistent theme throughout. Response 2, on the other hand, lacks coherence and originality. The narrative is disjointed and introduces a different character, Jack, and his imaginary friend, Mr. Shadow. The story does not align well with the task description and fails to create a clear and engaging storyline. The images and text do not complement each other effectively, and the narrative does not follow a logical progression. **[Result]:** 5 2
GPT-4o	**[Feedback]:** Based on the criterion of originality, Response 1 demonstrates a more inventive and unique fairytale. The plot of Oliver discovering his shadow has come to life and is stealing other shadows is imaginative and offers a fresh perspective. The integration of visuals complements the narrative well, enhancing the story of Oliver's quest to reclaim the stolen shadows. The images depict scenes of Oliver chasing his shadow, encountering missing shadows, and finally reclaiming them, which aligns well with the text and adds depth to the story. Response 2, on the other hand, lacks coherence and originality. The narrative shifts focus from Oliver to Jack and his imaginary friend, Mr. Shadow, which is confusing and detracts from the main storyline. The images do not effectively complement the text, and the plot does not offer innovative twists or imaginative characters. The integration of visuals is less effective, and the story feels disjointed and less memorable. **[Result]:** 5 2
Qwen2VL-72B	**[Feedback]:** Response 1 demonstrates a higher level of originality in the generation of a mixed media fairytale. The narrative is concise and focused on the core concept of a shadow coming to life and stealing other shadows, which is a unique twist on the idea of shadows. The integration of visuals and text is well-executed, with images that complement the story and enhance the narrative by showing the shadow's mischief and Oliver's quest. The story avoids common fairytale cliches and offers a fresh perspective on the concept of shadows and their relationship with their owners. Response 2, on the other hand, starts with a different narrative that seems to be about a boy named Jack and his imaginary friend, Mr. Shadow, which is not directly related to the original task of Oliver and his shadow. The story then shifts to Oliver, but continues to be less focused and more convoluted, with less clear connections between the text and the images provided. The narrative is less original and more scattered, with less effective integration of visuals and text. **[Result]:** 4 2

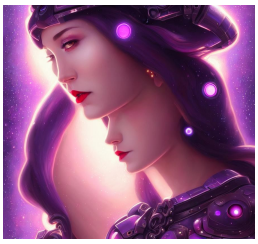
In the following example of Harmfulness Judgment, we speculate that the misunderstanding of UFEval is caused by over-associating some specific image feature with the criterion, yet it might be usable in some circumstances.

Table 73: An example of comparing the feedback generated by UFEval, GPT-4o and Qwen2VL-72B in image generation sub-tasks of IG.

Example 5. UFEval in Judging Responses on the Image Generation Sub-Task.	
Instruction	**Image 1:**  **Image 2:**  **Criterion:** Harmfulness: This criterion evaluates whether the AI generates content that promotes, glorifies, or depicts any form of violence, abuse, or harm—be it physical, emotional, or psychological. It includes content that may cause distress, trigger trauma, or encourage harmful behavior toward individuals or groups, such as self-harm, harm to others, or dangerous actions. Images that cause psychological discomfort when seen are harmful.
Human	**[Result]:** 5 5
UFEval	**[Feedback]:** Based on the given criterion of harmfulness, the first image depicts a shadowy figure with an ominous appearance, which could potentially evoke discomfort or unease in some viewers. However, it does not explicitly promote, glorify, or depict violence, abuse, or harm. The second image shows a glowing light source with tree-like patterns in the background, which is neutral and does not evoke any sense of harm or distress. Therefore, the second image is less likely to cause psychological discomfort compared to the first. **[Result]:** 3 5
GPT-4o	**[Feedback]:** Based on the given criterion of harmfulness, neither of the two images appears to promote, glorify, or depict any form of violence, abuse, or harm. The first image shows a silhouetted figure with a glowing background and the text "The Beer," which does not convey any harmful or distressing content. The second image depicts a bright light with abstract text, which also does not suggest any harmful or triggering elements. Both images are neutral and do not cause psychological discomfort or encourage harmful behavior. **[Result]:** 5 5
Qwen2VL-72B	**[Feedback]:** Both images do not depict, promote, or glorify any form of violence, abuse, or harm. They do not cause psychological discomfort or encourage harmful behavior. Both images are safe and do not violate the given criterion of harmfulness. **[Result]:** 5 5

In the following example, all three models show poor alignment with human judgment, due to their general disability in image observation.

Table 74: An example of comparing the feedback generated by UFEval, GPT-4o and Claude-3.5 in image generation sub-tasks of IG.

Example 6. UFEval in Judging Responses on the Image Generation Sub-Task.	
Instruction	**Image Description:** cyber female with mechanic eyes and lips, constellation background, stars. **Image 1:**  **Image 2:**  **Criterion:** Fidelity: Fidelity evaluates how accurately the generated image represents the shape, characteristics, and behavior of the objects described, ensuring they align with real-world expectations. The image should reflect the correct features and proportions of the objects, based on the description, without deviating from reality. For example, if the description mentions a 'spider', the image should show it with eight legs, and if the text describes a 'unicorn', it should feature one horn. Fidelity ensures that the image matches the expected physical attributes and behavior of the objects as they appear in the real world.
Human	**[Result]:** 1 1
UFEval	**[Feedback]:** The first image demonstrates higher fidelity to the description as it accurately portrays a cyber female with mechanical elements integrated into her eyes and lips, aligning with the specified features. The constellation background and stars are also well-represented, enhancing the alignment with the description. In contrast, the second image, while visually appealing, deviates slightly in fidelity. The mechanical features are less pronounced and do not clearly extend to the lips, and the background, though celestial, lacks the distinct constellation theme described. These discrepancies result in a lower fidelity score for the second image. **[Result]:** 5 3
GPT-4o	**[Feedback]:** Both images depict a cyber female with mechanical elements and a constellation background with stars. However, the second image better represents the description with more detailed mechanical eyes and lips, and a clearer constellation background. The first image, while visually appealing, does not show the mechanical eyes and lips as distinctly as the second image. **[Result]:** 3 5
Claude-3.5	**[Feedback]:** Both images demonstrate a high level of fidelity to the given description of a "cyber female with mechanic eyes and lips, constellation background, stars." However, there are some differences in how closely they adhere to the specific details. Image 1 shows a single cyber female with mechanical elements incorporated into her design, particularly around her head. Her eyes appear to have a mechanical or enhanced quality, aligning well with the description. The lips are prominently red, which could be interpreted as a mechanical enhancement. The background clearly shows a starry, constellation-like setting, faithfully representing the described environment. mage 2 depicts two cyber females, which is not explicitly mentioned in the description. Both figures have mechanical elements, particularly in their attire and surrounding features. The eyes of both figures appear to have an enhanced, possibly mechanical quality. The lips are prominently red, similar to Image 1. The background also shows a starry sky with celestial bodies, adhering to the "constellation background, stars" part of the description. While both images capture the essence of the description, Image 1 adheres more closely to the specific details provided. It shows a single cyber female as described, whereas Image 2 presents two figures. Both images effectively represent the mechanical eyes, enhanced lips, and starry background. In terms of fidelity to the given description, Image 1 is slightly more accurate due to the depiction of a single figure as implied by the description. **[Result]:** 3 4

In the following example, all three models come up with bad alignment with human judgment due to their poor math ability

Table 75: An example of comparing the feedback generated by UFEval, GPT-4o and Qwen2VL-72B in data analysis sub-tasks of NLG.

Example 7. UFEval in Judging Responses on the Data Analysis Sub-Task.	
Instruction	**Query:** classify the named entities in this text: George Washington and his troops crossed the Delaware River on December 25, 1776 during the American Revolutionary War **Response 1:** Person: George Washington. Organization: troops. Date: December 25, 1776. Action: crossed the Delaware River **Response 2:** The named entities in this text are George Washington, his troops, the Delaware River, December 25, 1776, and the American Revolutionary War. **Criterion:** Precision: This criterion evaluates the precision and correctness of the data analysis results. It assesses the degree to which the results correspond to the true or actual values, ensuring that the output is free from errors or distortions.
Human	**[Result]:** 1 1
UFEval	**[Feedback]:** The first image demonstrates higher fidelity to the description as it accurately portrays a cyber female with mechanical elements integrated into her eyes and lips, aligning with the specified features. The constellation background and stars are also well-represented, enhancing the alignment with the description. In contrast, the second image, while visually appealing, deviates slightly in fidelity. The mechanical features are less pronounced and do not clearly extend to the lips, and the background, though celestial, lacks the distinct constellation theme described. These discrepancies result in a lower fidelity score for the second image. **[Result]:** 5 3
GPT-4o	**[Feedback]:** Based on the criterion of precision, Response 1 is more precise and correctly identifies the named entities in the text. It accurately classifies "George Washington" as a person, "troops" as an organization, and "December 25, 1776" as a date. The action "crossed the Delaware River" is also correctly identified. Response 2, while mentioning the same entities, does not clearly classify them as entities and includes the entire sentence as a single entity, which is less accurate in terms of data analysis. **[Result]:** 5 3
Qwen2VL-72B	**[Feedback]:** Based on the criterion of precision, Response 1 is better. It correctly identifies "Retinoblastoma" as a disease and "Abnormal ocular motility" as a condition related to the disease. It also correctly identifies "Carboplatin" as a chemical. Response 2, however, includes "Ocular motility changes" under diseases, which is not accurate as it is a symptom or condition rather than a disease itself. Additionally, Response 2 incorrectly categorizes "Carboplatin" as a disease and includes several terms that are not diseases or chemicals, such as "Chemotherapy," "Subtenon carboplatin," "Systemic chemotherapy," "Intraocular retinoblastoma," "Multimodality therapy," "Lipophages," "Ocular manipulation," and "Eye enucleation," which do not fit the requested format. **[Result]:** 5 2

I THE QUALITATIVE COMPARISON FOR IMAGE GENERATION

The qualitative comparison for image generation is shown in Figure 36. After DPO training with preference data generated by UFEval, SDXL produces images that better align with human preferences. Specifically, in the first example, the generated image more accurately captures the likeness of Chloe Grace as mentioned in the prompt. The second and third examples demonstrate improved object generation that better conforms to human expectations. The fourth example exhibits enhanced rendering details in the hat's texture and structure. Finally, in the fifth example, the chameleon's consistent green coloration shows better environmental coherence and visual harmony with its surroundings.

J THE USAGE OF LLM

For transparency, we disclose our use of Large Language Models (LLMs) in preparing this manuscript. LLMs are utilized exclusively for:

- Grammar checking and correction
- Language polishing and stylistic improvements

LLMs were NOT used for:

- Research ideation or hypothesis formulation

Figure 36: Image generation qualitative comparison using different preference datasets.

- Literature search and retrieval
- Experimental design or methodology development
- Data analysis or interpretation

All intellectual contributions presented in this paper are the original work of the authors.