Multi-modal, Multi-task, Multi-criteria Automatic Evaluation with Vision Language Models

Masanari Oi [1], **Masahiro Kaneko** [2,1], **Naoaki Okazaki** [1,3], **Nakamasa Inoue** [1]
[1] Institute of Science Tokyo, [2] MBZUAI, [3] NII LLMC
ohi.m.7b5f@m.isct.ac.jp

Abstract

Vision-language models (VLMs) have shown impressive abilities across a range of multi-modal tasks. However, existing metrics for evaluating the quality of text generated by VLMs typically focus on an overall evaluation for a specific task, such as image captioning. While the overall evaluation is essential for any task, the criteria prioritized can differ depending on the task, making it challenging for current metrics to adapt to multi-task scenarios. To address this limitation, we propose HarmonicEval, a reference-free comprehensive evaluation metric that aggregates criterion-wise scores to produce the overall score in a bottom-up manner. Furthermore, to assess the generalizability of automatic evaluation metrics in multi-task scenarios, we construct the Multi-task Multi-criteria Human Evaluation (MMHE) benchmark, which comprises 18,000 expert human judgments across four multi-modal tasks. Our experiments demonstrate that HarmonicEval achieves higher correlations with human judgments than conventional metrics while providing numerical scores for each criterion. Project page: https://stjohn2007.github.io/MMHE_project/

Keywords: Automatic Evaluation, Annotated Dataset, Vision-Language models

1. Introduction

Automatic evaluation of text generated by vision-language models (VLMs) is essential for improving their performance across various multi-modal tasks, such as image captioning and visual question answering (Hessel et al., 2021; Lee et al., 2024b). As the range of tasks that VLMs can perform continues to expand, developing specialized evaluation metrics for each task becomes increasingly difficult. Hence, a comprehensive metric capable of evaluating text across multiple tasks is highly desirable. However, most existing metrics focus on measuring the overall quality of text within a specific task, limiting their applicability in multi-task settings (see Figure 1 (a)).

Existing metrics that provide only overall scores (Papineni et al., 2002; Zhang et al., 2020; Hessel et al., 2021) often prioritize specific evaluation criteria, as discussed in previous studies (Kasai et al., 2022; Fabbri et al., 2021). For example, metrics for evaluating image captions typically prioritize correctness and completeness over conciseness and fluency. When these metrics are applied to other tasks, such as visual question answering, they tend to overvalue verbose or unnatural responses. To address this limitation, integrating multiple evaluation criteria to predict the overall score, a concept we refer to as *comprehensive evaluation*, holds significant potential for a more comprehensive assessment in multi-task scenarios. However, this approach remains underexplored due to the lack of a meta-evaluation benchmark that provides human judgment across multiple tasks and criteria.

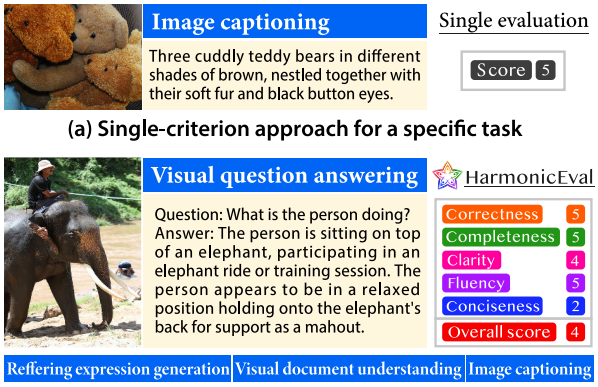

Figure 1: **Multi-task and multi-criteria evaluation.** (a) Conventional single-criterion approach focuses on a single task, such as image captioning. (b) HarmonicEval integrates multiple criteria to provide overall scores. MMHE consists of 18,000 expert human judgments across four multi-modal tasks and five criteria.

This motivates us to introduce a novel evaluation metric with a new benchmark in this paper.

First, we propose HarmonicEval, a reference-free harmonic evaluation metric for multiple multi-modal tasks. As shown in Figure 1 (b), HarmonicEval integrates multiple criteria to produce the overall score in a bottom-up manner. Specifically, the evaluation pipeline consists of two steps: 1) criterion-wise scoring, where a VLM is prompted to evaluate the input text based on each specific criterion, and 2) score aggregation, where the overall score is calculated from the criterion-wise scores. For score

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9396–9408
11–16 May 2026.

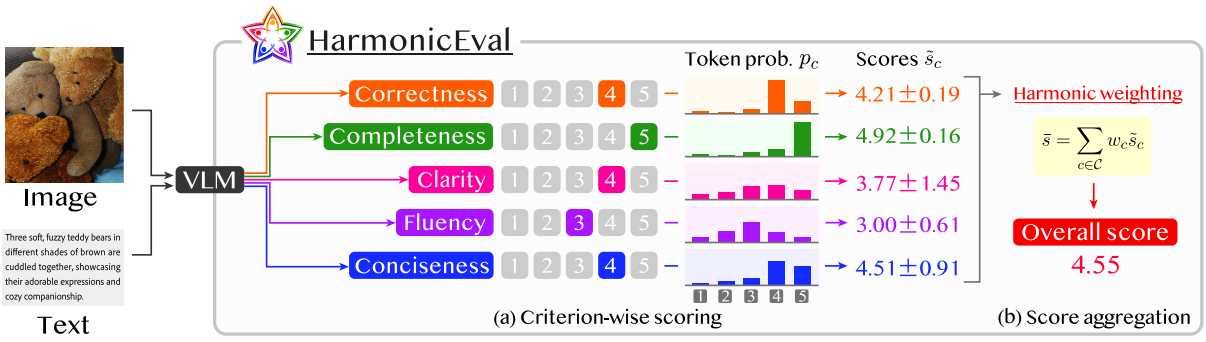

Figure 2: HarmonicEval framework consists of two steps. (a) Criterion-wise scoring is performed by prompting a VLM to evaluate the input text based on each criterion, followed by score smoothing to improve robustness based on the first-order statistics. (b) Score aggregation produces an overall score using harmonic weighting based on the second-order statistics, aiming to reduce statistical fluctuations.

aggregation, we introduce a novel harmonic weighting scheme that automatically determines weight coefficients based on the second-order statistics of the output token probability distributions.

Second, to assess the generalizability of automatic evaluation metrics, we introduce the multi-task multi-criteria human evaluation (MMHE) benchmark, the first meta-evaluation benchmark that provides human judgment annotations across multiple criteria and multiple multi-modal tasks. Specifically, MMHE consists of 18,000 expert human judgments on five criteria for four diverse tasks: referring expression generation (REG), visual question answering (VQA), visual document understanding (VDU), and image captioning (IC).

Our experiments on MMHE show that HarmonicEval achieves higher correlations with human judgments than conventional metrics, while also providing criterion-specific scores that highlight areas for improvement. Furthermore, we demonstrate that HarmonicEval achieves state-of-the-art or comparable performance in conventional image caption evaluation scenarios across five widely used benchmarks that provide only overall judgments: Flickr8k-EX / CF (Hodosh et al., 2013), Composite (Aditya et al., 2018), PASCAL-50S (Vedantam et al., 2015), and FOIL (Shekhar et al., 2017). In summary, our key contributions are threefold:

- We propose HarmonicEval, a novel reference-free metric for harmonic evaluation across multiple multi-modal tasks.
- We introduce MMHE, the first multi-task multi-criteria human evaluation benchmark, consisting of 18,000 expert judgments spanning four multi-modal tasks and five evaluation criteria.
- We demonstrate the effectiveness of HarmonicEval on MMHE and five conventional image captioning benchmarks. In addition, we conduct the first in-depth analysis of how existing metrics implicitly prioritize different evaluation criteria.

2. HarmonicEval

As shown in Figure 2, the pipeline of HarmonicEval consists of two steps: criterion-wise scoring (§2.1) and score aggregation (§2.2).

2.1. Criterion-wise scoring

In this step, a VLM is employed as an evaluator and prompted to generate evaluation scores on each criterion independently. Let $\boldsymbol{t}$ be an input text to be evaluated, such as an image caption for the IC task. The evaluation process to obtain criterion-wise scores s_c is formulated as $s_c = f([\boldsymbol{p}_c, \boldsymbol{t}], \boldsymbol{x})$, where c is a criterion, $\boldsymbol{x}$ is an input image, f is a VLM, $\boldsymbol{p}_c$ is a prompt, and $[,]$ denotes textual concatenation. To improve alignment with human judgments, score smoothing (Liu et al., 2023b; Lee et al., 2024b) is applied as $\tilde{s}_c = \sum_{r \in R} rP(r \mid [\boldsymbol{p}_c, \boldsymbol{t}], \boldsymbol{x})$ where $P(r \mid [\boldsymbol{p}_c, \boldsymbol{t}], \boldsymbol{x})$ is the output token probability of token r assigned by the VLM, and R is a set of ratings.

2.2. Score aggregation

To aggregate the criterion-wise scores $\tilde{s}_c$, we introduce harmonic weighting, a novel approach that leverages the second-order statistics of the output token probability distributions to adaptively determine the weight coefficients for aggregation. Compared to simple averaging, our aggregation approach aims to better align with human evaluation by dynamically emphasizing more reliable scores based on the input. Specifically, the overall score S is computed as

$$S = \sum_{c \in \mathcal{C}} w_c \tilde{s}_c, \quad w_c = \frac{1}{H} \sigma_c^{-2(1-\gamma)/\gamma}, \qquad (1)$$

where w_c is a weight coefficient, $\mathcal{C}$ is a set of criteria, σ_c is the standard deviation of the criterion-wise score given by

$$\sigma_c = \sqrt{\sum_{r \in R} (r - \tilde{s}_c)^2 \; P(r \mid [\boldsymbol{p}_c, \boldsymbol{t}], \boldsymbol{x})}, \qquad (2)$$

and $H = \sum_{c \in \mathcal{C}} \sigma_c^{-2(1-\gamma)/\gamma}$ is the harmonic mean of the variances with a hyperparameter γ. Smaller values of σ_c can be interpreted as indicating higher confidence in the evaluation of c. The role of hyperparameter γ is to bridge three weighting strategies: uniform weighting, inverse variance weighting and selective weighting as detailed below.

Uniform weighting. When $\gamma = 1$, harmonic weighting reduces to uniform weighting $w_c = 1/|\mathcal{C}|$. This is effective when all criterion-wise scores are equally reliable in determining the overall score. However, this does not provide the best estimator as observed variances are ignored in aggregation.

Inverse variance weighting. When $\gamma = 0.5$, harmonic weighting reduces to inverse variance weighting $w_c \propto \sigma_c^{-2}$. This provides the best linear unbiased estimator under the assumption that the observed variance is due to statistical fluctuations. However, this assumption is not always reasonable, as each criterion may have its own variance.

Selective weighting. When $\gamma \to 0$, only the score $\tilde{s}_c$ with the smallest variance is selected as the overall score. This approach is used in experiments to show the necessity of aggregation.

Discussion. When the evaluation criteria are carefully designed, the uniform weighting ($\gamma = 1.0$) aligns closely with human expert judgment, and $0.5 \leq \gamma < 1.0$ further improves the alignment because it adaptively reflects the confidence of criterion-wise scores. Since the assumption underlying the inverse variance weighting ($\gamma = 0.5$) is not reasonable when each criterion has its own variance, we hypothesize that a value between 0.5 and 1.0 is optimal and choose $\gamma = 0.75$ as the default value.

2.3. Implementation details

Definition of criteria. Based on prior research in natural language generation (Asano et al., 2017; Kryscinski et al., 2019; Fabbri et al., 2021; Freitag et al., 2021; Song et al., 2024) and multi-modal evaluation (Aditya et al., 2018; Kasai et al., 2022), we define five evaluation criteria: $\mathcal{C} = \{\text{Correctness}, \text{Completeness}, \text{Fluency}, \text{Conciseness}, \text{Clarity}\}$. Their definitions are summarized in Table 1. We validated these criteria by examining 100 outputs from various vision-language tasks and confirmed their adequacy for reliable evaluation.

Correctness (Crt): The degree to which the target text accurately reflects the content of the input image and text.
Completeness (Cmp): The extent to which the target text captures all relevant and significant details of the input image and text.
Clarity (Clr): The ease with which the reader can understand the target text.
Fluency (Flu): The grammatical accuracy and natural flow of the target text.
Conciseness (Cnc): The efficiency of the target text in conveying information without unnecessary verbosity.

Table 1: Five criteria for HarmonicEval and MMHE.

Aggregating these five criteria contributes to the overall text quality evaluation across a wide range of tasks. Depending on the task and the style of the input text, some criteria may not be necessary. However, when such evaluations are conducted, the output becomes less confident, leading to higher variance σ_c and lower weight coefficients w_c, as VLMs account for task and criterion features in addition to the input text. Thus, HarmonicEval can adaptively perform comprehensive evaluations without manual tuning of weight coefficients.

Prompts. The prompt $\boldsymbol{p}_c$ instructs the VLM to evaluate text with respect to the criterion c on a five-point scale $R = \{1, 2, 3, 4, 5\}$. Below is the prompt of the correctness criterion for the IC task.

*Your task is to rate the **caption** for the given image on a scale of 1 to 5 on the following criterion and rating scale.*
Evaluation Criterion: Correctness
*How accurately does **the caption describe the image?***
Rating Scale:
*- 1 Very Low Correctness: The **caption** is mostly or entirely incorrect ...*
...
*- 5 Extremely High Correctness: The **caption** perfectly captures all ...*

Here, the boldfaced portions indicate the task-dependent phrases. For example, "caption" is replaced with "answer" for the VQA task.

3. MMHE Benchmark

We present the MMHE benchmark, the first meta-evaluation benchmark that covers multiple evaluation criteria across multiple multi-modal tasks. We collected 18,000 expert human judgments spanning four tasks and five criteria. Example human judgment scores are shown in Figure 3.

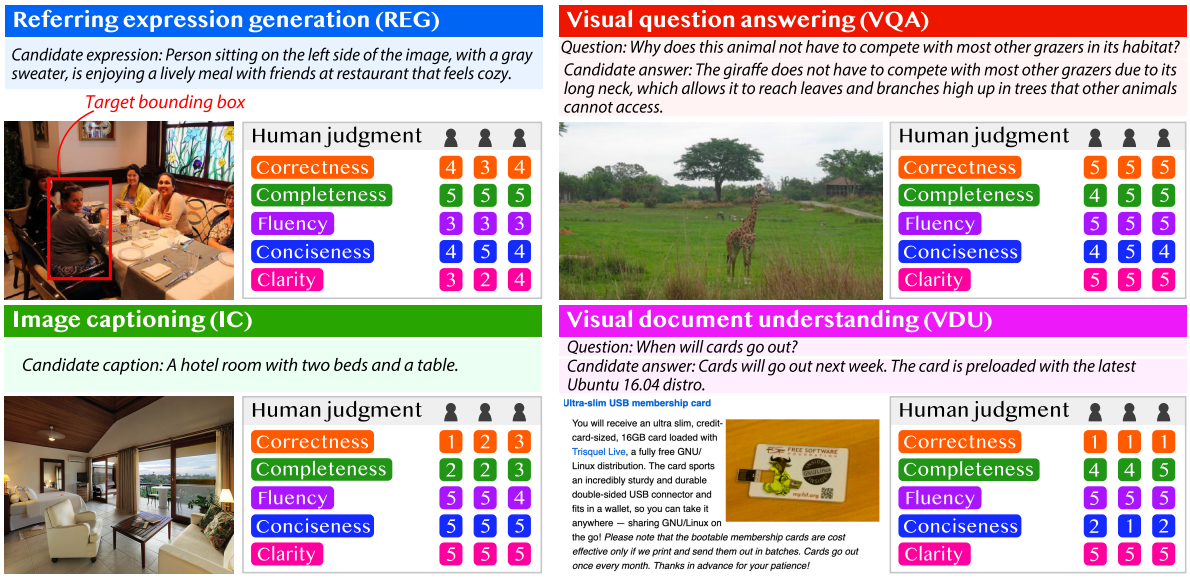

Figure 3: MMHE benchmark is a multi-task multi-criteria human evaluation benchmark. Each candidate text is manually evaluated by three expert annotators.

3.1. Benchmark design

Motivation. Despite numerous human evaluation benchmarks for multi-modal tasks, most focus on a single task (e.g., image captioning) or offer only an overall quality label. Our purpose is twofold: 1) to assess how evaluation metrics perform in a multi-task setting, and 2) to investigate how metrics that provide only overall scores prioritize certain criteria over others. To achieve these goals, we design MMHE to cover multiple multi-modal tasks and to explicitly collect human judgments across five evaluation criteria.

Multi-modal tasks. We select four diverse tasks to show how different criteria matter across contexts: 1) **REG** aims to generate a textual expression that uniquely identifies a specific object in the image (marked by a bounding box). We expect completeness to be crucial for precisely distinguishing the target object. 2) **VQA** requires generating an answer to a question about the image content. Given the nature of question answering, we hypothesize that correctness and conciseness are particularly important here. 3) **VDU** focuses on interpreting information from visually presented documents. Similar to VQA, we suspect that correctness and conciseness play key roles. 4) **IC** involves producing a descriptive sentence for the entire image. We anticipate that correctness and completeness are especially relevant for capturing key elements.

MMHE is the first benchmark to integrate multiple multi-modal tasks with a unified set of evaluation criteria in a single framework. This design enables fine-grained, criterion-wise analysis of how different metrics perform across various task requirements, which cannot be achieved by simply combining existing benchmarks with their disparate evaluation criteria.

3.2. Benchmark construction

The benchmark construction process consists of three steps: 1) Source selection, which selects source text-image pairs; 2) Target generation, which creates target texts to be evaluated using state-of-the-art VLMs; and 3) Human expert evaluation, which assesses the quality of the target texts.

Source selection. We selected the following four datasets: RefCOCO (Kazemzadeh et al., 2014) for REG, OK-VQA (Marino et al., 2019) for VQA, VisualMRC (Tanaka et al., 2021) for VDU, and MSCOCO (Lin et al., 2014) for IC. We randomly sampled 100 instances from the validation or test subset of each dataset.

Target generation. The target texts to be evaluated were generated using state-of-the-art VLMs. Specifically, we employed ten VLMs: LLaVA-1.5-7B/13B (Liu et al., 2023a), InstructBLIP-Vicuna-7B/13B (Dai et al., 2023), Qwen-VL (Bai et al., 2023), Qwen2-VL-Instruct-7B/72B (Wang et al., 2024a), CogVLM-Chat (Wang et al., 2024b), GPT-4o-mini and GPT-4o (OpenAI, 2024). For each instance, we assigned three distinct VLMs from this pool, ensuring that every instance had exactly three candidate outputs. The assignment was balanced across tasks and models. This design resulted in 100 instances × 3 outputs = 300 candidate responses per task.

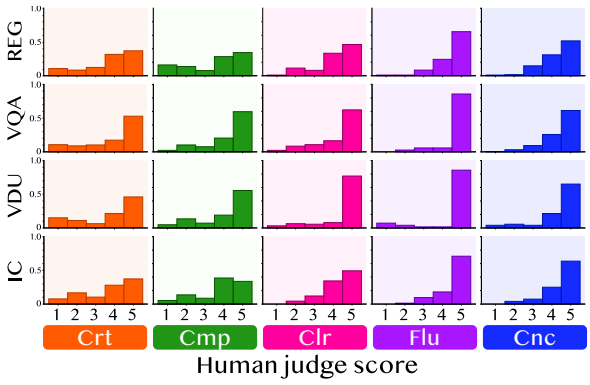

Figure 4: Human judgment score distributions for each task and criterion on the MMHE benchmark.

Human expert evaluation. Five expert annotators were given an explanation of the four multimodal tasks and asked to carefully review the five evaluation criteria in Table 1. They rated each target text on a five-point scale, then conducted a thorough review for consistency. Each instance was independently scored by three annotators, yielding 18,000 human judgment scores.

Overall judgments. To collect overall judgments, we adopted a best-of-three approach, in which annotators choose the best output among three responses generated by different models for the same input. We did not use a five-point scale for the overall score because defining each rating level (from 1 to 5) for the overall quality is difficult and could introduce bias (Chiang et al., 2024).

3.3. Data analysis

Figure 4 presents the score distributions for each criterion across the four tasks. The correctness and completeness criteria exhibit diverse score distributions for most tasks, suggesting that even the state-of-the-art VLMs face challenges in these aspects. In contrast, the fluency criterion shows a narrow score distribution across most tasks, with a dominant score of five. This indicates that the VLMs generate fluent text even when visual understanding is inaccurate. Nonetheless, we consider fluency an essential criterion, as overlooking it could lead to high scores being assigned to correct but non-fluent text. A qualitative example illustrating this is provided in Section 4.2. The clarity and conciseness criteria have score distributions that are intermediate between the diverse and narrow distributions. While the score of five is more prevalent than in the diverse distributions, other scores still occur with noticeable frequency.

Method	REG	VQA	VDU	IC	Avg.
BLEU	45.3	29.4	57.3	46.8	44.7
ROUGE	49.0	30.8	56.0	47.9	45.9
CIDEr	42.5	25.0	62.1	42.7	43.1
METEOR	44.4	29.4	59.7	53.6	46.8
BERT-S	46.2	33.8	62.1	53.1	48.8
BART-S	56.4	20.5	60.9	57.8	48.9
CLIP-S	60.1	39.7	60.9	52.0	53.2
G-VEval	60.1	75.0	71.9	68.7	68.9
FLEUR	62.9	**76.4**	60.9	73.9	68.5
GPT-FLEUR	60.1	75.0	**76.5**	76.0	71.9
HarmonicEval	**66.6**	**76.4**	73.4	**77.0**	**73.4**

Table 2: Accuracy (%) on MMHE. The best result for each task is marked in bold. Average (Avg.) indicates the average accuracy across the four tasks.

4. Experiments

4.1. Performance on MMHE

We evaluate the performance of HarmonicEval on MMHE and compare it with conventional metrics. We also examine how existing metrics prioritize or deprioritize specific criteria in each task.

Settings. To assess the performance of each metric, we use accuracy (%) for the overall evaluation and the Kendall's tau correlation coefficient τ for criterion-wise evaluations.

We implement nine baselines, including four n-gram-based metrics (Papineni et al., 2002; Lin, 2004; Vedantam et al., 2015; Banerjee and Lavie, 2005) and five neural network-based metrics (Zhang et al., 2020; Yuan et al., 2021; Hessel et al., 2021; Tong et al., 2025; Lee et al., 2024b), grouped in Table 2. Among them, FLEUR (Lee et al., 2024b) is a state-of-the-art VLM-based metric[1]. HarmonicEval utilizes GPT-4o as its backbone. For a fair comparison with FLEUR, we also implement GPT-FLEUR, which substitutes GPT-4o for the original LLaVA-1.5-13B.

Main results. Table 2 compares HarmonicEval with conventional metrics in terms of overall performance on MMHE. HarmonicEval achieves the highest accuracy in REG (66.6), VQA (76.4), IC (77.0), and attains the top average score of 73.4 across tasks. While GPT-FLEUR obtains the highest score on VDU (76.5), it performs less effectively on REG. These results underscore the strong multi-task capability of HarmonicEval.

Correlation analysis. To investigate how existing metrics prioritize or deprioritize certain criteria, we

[1]Note that all baseline metrics don't support criterion-wise scoring and produce only overall scores.

Metric	REG					VQA					VDU					IC				
	Crt	Cmp	Clr	Flu	Cnc	Crt	Cmp	Clr	Flu	Cnc	Crt	Cmp	Clr	Flu	Cnc	Crt	Cmp	Clr	Flu	Cnc
BLEU	6.0	6.9	3.9	1.2	6.1	-1.3	-10.4	-11.0	-19.3	4.1	19.8	12.9	14.9	14.3	21.2	4.4	4.5	5.9	0.3	11.3
ROUGE	2.3	5.7	4.4	-3.5	3.9	7.1	-2.8	-5.0	-8.1	10.2	20.0	14.7	16.2	17.9	22.7	5.2	6.5	9.0	4.4	9.7
CIDEr	6.4	3.4	2.4	-9.7	20.9	-27.8	-39.0	-19.5	-26.0	-3.8	23.7	15.8	19.3	18.0	23.8	0.7	-1.6	8.7	-3.8	14.5
METEOR	1.9	5.3	5.2	-5.1	-6.3	5.3	-3.9	-8.2	-8.5	2.7	17.8	18.0	16.9	20.5	14.9	6.8	12.1	7.3	-2.3	1.0
BERT-S	6.5	6.9	-6.5	-8.6	12.4	-2.8	-14.3	4.9	-10.0	6.1	21.0	17.4	20.4	21.6	23.9	12.3	11.1	6.4	4.7	10.5
BART-S	4.4	6.7	4.2	-7.8	3.1	-13.4	-20.2	-2.8	-16.6	1.6	22.4	21.3	21.6	17.9	14.7	4.8	4.3	4.3	2.2	3.2
CLIP-S	13.5	14.4	6.8	-0.9	-5.1	6.6	5.4	7.2	8.1	4.5	15.2	12.5	15.0	12.6	8.4	20.2	21.3	11.1	3.2	3.5
G-VEval	11.4	23.4	18.7	9.9	8.3	52.8	41.0	19.0	44.7	35.9	54.1	41.7	47.0	40.6	42.0	43.8	43.4	21.9	26.0	14.5
FLEUR	**29.3**	**30.8**	18.6	8.7	11.2	38.7	38.2	**39.9**	39.8	**44.7**	38.1	37.1	44.6	35.2	28.2	33.9	35.0	**25.9**	24.5	14.0
GPT-FLEUR	19.7	30.6	14.0	19.7	11.0	**54.7**	42.0	18.0	35.0	23.0	59.0	44.4	43.8	37.2	29.5	**47.5**	47.7	25.3	29.1	13.4
HarmonicEval	23.2	**30.8**	**24.0**	**20.7**	**23.8**	53.5	**50.6**	31.8	**51.9**	44.4	**60.0**	**48.8**	**47.9**	**51.2**	**45.8**	44.7	**50.3**	19.8	**36.4**	**22.8**

Table 3: Criterion-wise correlation analysis on MMHE. Scores for the most positively and negatively correlated criteria are marked with red and blue underlines, respectively. The highest correlations for each criterion are highlighted in bold. Crt: Correctness, Cmp: Completeness, Clr: Clarity, Flu: Fluency, Cnc: Conciseness.

show the correlation between the predicted overall scores and the human judgment scores for each criterion in Table 3. In the table, red and blue underlines denote the most and least correlated criterion, respectively, for each task.

We observe task-wise trends. For REG, completeness shows the highest correlation across most metrics. This is reasonable, as REG requires explicit expressions to identify a unique object by distinguishing it from marked objects. For VQA, most metrics are more strongly correlated with conciseness but less so with completeness. This indicates that conventional metrics deprioritize completeness, potentially leading to inaccurate evaluations of insufficient answers. A similar trend is observed in VDU, where completeness is also deprioritized by conventional metrics. For IC, fluency exhibits low correlations for most metrics, suggesting a tendency to assign high scores even to non-fluent texts. Overall, these results underscore the necessity of a comprehensive evaluation metric in multi-task scenarios.

Table 3 also shows the correlations between the criterion-wise scores of HarmonicEval and the human judgment scores. HarmonicEval achieves the highest correlation across most criteria. While this is expected, as HarmonicEval is the only metric that outputs criterion-wise scores, the result nevertheless demonstrates that its predictions align well with human judgments on each individual criterion.

4.2. Analysis

Can HarmonicEval improve explainability? Providing feedback on evaluation results to users is important. To investigate whether HarmonicEval offers clear textual explanations, we prompt the VLM with "*Why? Tell me the reason.*" after obtaining the overall scores, following Lee et al. (2024b).

	REG	VQA	VDU	IC	Total
HarmonicEval	**19***	**12**	**21***	**19***	**71***
FLEUR	3	9	3	3	18
Tie	3	4	1	3	11

Table 4: User study on textual explainability. Asterisks ($*$) denote statistical significance between HarmonicEval and FLEUR ($p < 0.05$, binomial test).

Figure 5 shows a qualitative comparison of HarmonicEval and FLEUR on the IC task. Panel (a) shows that HarmonicEval successfully detects severe fluency issues and reflects them in its overall score, whereas FLEUR tends to overlook such deficiencies. This suggests that FLEUR may fail to capture certain criteria (e.g., fluency), as discussed in Section 3.3. Panel (b) illustrates that HarmonicEval identifies incorrect details in captions more precisely than FLEUR, assigning a lower correctness score accompanied by a criterion-specific explanation. These examples indicate that HarmonicEval provides more informative textual feedback.

To verify this observation more systematically, we conduct a user study to quantitatively assess the explainability of HarmonicEval. Specifically, we sample 25 instances from each task and generate explanations for both HarmonicEval and FLEUR, yielding 100 explanation pairs in total. Five human annotators then evaluate which explanation in each pair is more informative and useful, allowing for ties. As shown in Table 4, HarmonicEval significantly outperforms FLEUR on textual explainability, consistent with its more fine-grained criterion-wise evaluation observed in the qualitative examples.

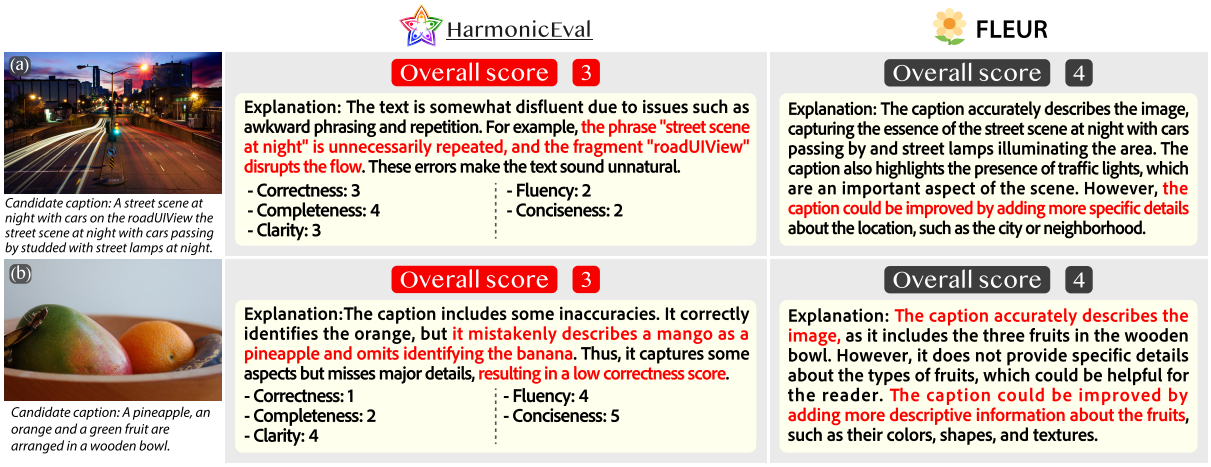

Figure 5: Qualitative examples.

Metric	REG	VQA	VDU	IC	Avg.
HarmonicEval	**66.6**	**76.4**	73.4	**77.0**	**73.4**
w/o criterion-wise scoring	62.0	73.5	**75.9**	76.5	72.0
w/o harmonic weighting	65.7	75.0	73.4	76.5	72.6

Table 5: Ablation study.

γ	REG	VQA	VDU	IC	Avg.
0.01	47.2	69.1	61.4	53.1	57.7
0.50	**66.6**	**76.4**	**73.4**	76.5	73.2
0.75	**66.6**	**76.4**	**73.4**	**77.0**	**73.4**
1.00	65.7	75.0	**73.4**	76.5	72.6

Table 6: Hyperparameter study.

Is each component in HarmonicEval essential? We conduct an ablation study on the two key components of HarmonicEval: criterion-wise scoring and harmonic weighting. Specifically, we examine two alternative approaches: 1) prompting the VLM to directly predict an overall score based on all five criteria, without using criterion-wise scores (*w/o criterion-wise scoring*); and 2) computing the overall score as a simple average of the criterion-wise scores (*w/o harmonic weighting*).

Table 5 shows that both components improve overall performance. Removing criterion-wise scoring lowers scores on REG, VQA, and IC, indicating that explicitly scoring each criterion results in better evaluation than relying on a single overall score based on a detailed prompt. Similarly, omitting harmonic weighting reduces performance in most tasks, validating the effectiveness of our approach.

Is statistical aggregation effective? Table 6 shows a hyperparameter study for γ. As expected, harmonic weighting with $\gamma = 0.75$ performs the best. This justifies the importance of the statistical aggregation process.

VLM	Method	REG	VQA	VDU	IC	Avg.
L-7B	FLEUR	**69.4**	69.1	**63.2**	72.6	**68.6**
	Harmonic	62.9	**72.0**	61.4	**73.4**	67.4
L-13B	FLEUR	62.9	76.4	60.9	**73.9**	68.5
	Harmonic	**64.8**	**77.9**	**63.2**	72.9	**69.7**
G-4o	FLEUR	60.1	75.0	**76.5**	76.0	71.9
	Harmonic	**66.6**	**76.4**	73.4	**77.0**	**73.4**

Table 7: Comparison of HarmonicEval and FLEUR across different backbone VLMs.

Is HarmonicEval effective across various backbone VLMs? Table 7 shows the overall performance comparison between HarmonicEval and FLEUR using LLaVA-1.5-7B, LLaVA-1.5-13B, and GPT-4o as backbone models. HarmonicEval consistently outperforms FLEUR on both LLaVA-1.5-13B and GPT-4o. On the other hand, FLEUR slightly outperforms HarmonicEval when using LLaVA-1.5-7B. We found that this is because LLaVA-1.5-7B tends to underrate texts in the conciseness criterion compared to human-assigned scores. Nonetheless, these results highlight HarmonicEval's effectiveness across different VLMs, particularly with more capable ones.

4.3. Performance on existing IC benchmarks

To assess the robustness of HarmonicEval in standard image captioning scenarios, we also evaluate it on five widely used IC benchmarks: Flickr8k-EX / CF, Composite, Pascal-50S, and FOIL. They support comparisons with a broader set of baselines, including specialized IC metrics (Anderson

	Metric	F-EX τ_c	F-CF τ_b	Com τ_c	Pas acc.
Reference-based	BLEU	30.8	16.9	30.6	72.9
	ROUGE	32.3	19.9	32.4	74.1
	METEOR	41.8	22.2	38.9	78.0
	CIDEr	43.9	24.6	37.7	76.8
	SPICE	44.9	24.4	40.3	69.6
	BERT-S	39.2	22.8	30.1	79.1
	TIGEr	49.3	–	45.4	80.7
	ViLBERTS-F	50.1	–	52.4	79.6
	FAIEr-4	52.6	35.4	57.7	81.4
	RefCLIP-Score	53.0	36.4	55.4	83.1
	Polos	56.4	37.8	57.6	**86.5**
	RefFLEUR	51.9	38.8	64.2	85.5
Reference-free	UMIC	46.8	30.1	56.1	85.1
	FAIEr-r	50.1	32.4	50.5	–
	CLIP-S	51.5	34.4	53.8	80.7
	InfoCLIP	32.6	23.5	15.3	64.1
	InfoMetIC	54.2	36.3	59.2	85.3
	InfoMetIC$^+$	55.5	36.6	59.3	**86.5**
	G-VEval	**59.7**	38.7	63.0	82.3
	FLEUR	53.0	38.6	63.5	83.2
	GPT-FLEUR	53.5	39.0	61.5	82.6
	HarmonicEval	53.1	**39.2**	**66.2**	82.4

Table 8: Comparison on Flickr8k-EX / CF (F-EX/CF), Composite (Com), and Pascal-50S (Pas).

Metric	1-ref	4-ref
Polos	93.3	95.4
RefFLEUR	97.3	**98.4**
FLEUR	96.8	96.8
GPT-FLEUR	97.0	97.0
HarmonicEval	**97.8**	97.8

Table 9: Comparison on FOIL.

et al., 2016; Jiang et al., 2019; Lee et al., 2020; Hu et al., 2023; Wada et al., 2024). Following prior works (Hessel et al., 2021; Lee et al., 2024b), we use Kendall's tau-c for Flickr8k-EX and Composite, tau-b for Flickr8k-CF, and accuracy for Pascal-50S.

As shown in Tables 8 and 9, HarmonicEval matches or surpasses state-of-the-art metrics on Flickr8k-CF, Composite, and FOIL. Despite modest gaps on Flickr8k-EX and Pascal-50S relative to computationally heavier or fine-tuned, task-specific metrics, strong results on the remaining three benchmarks underscore HarmonicEval's overall robustness without task-specific tuning.

5. Related Work

Automatic evaluation metrics. Traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) were developed for automatic evaluation of natural language generation (NLG), relying on n-gram overlap with references. Recently, large language models (LLMs) have been increasingly employed as evaluators across various NLG tasks (Kocmi and Federmann, 2023; Chiang and Lee, 2023; Zheng et al., 2023; Song et al., 2024). For example, G-Eval (Liu et al., 2023b) introduced a form-filling paradigm for criterion-based evaluation in summarization and dialogue generation. Our HarmonicEval extends this line by handling various multi-modal tasks and aggregating criterion-wise scores through second-order statistics of token probability distributions from VLMs.

For the IC task, CIDEr (Vedantam et al., 2015) measures the consensus between candidate and reference captions by weighting n-grams using TF-IDF. Recent metrics leverage VLMs to offer more flexible evaluation paradigms (Zhang et al., 2023; Lee et al., 2024a; Yu et al., 2024; Zhuang et al., 2024; Maeda et al., 2024; Chen et al., 2024; Tong et al., 2025). FLEUR (Lee et al., 2024b), a state-of-the-art reference-free metric, provides textual explanations that underlie its overall scores. However, existing metrics primarily focus on providing the overall evaluation, and often overlook specific criteria. HarmonicEval addresses these issues by offering criterion-wise scores alongside the overall score, enabling comprehensive evaluation.

Human evaluation benchmarks. Several human evaluation benchmarks have been proposed for NLG tasks. Example benchmarks include SummEval (Fabbri et al., 2021) for text summarization and the WMT shared tasks (Semenov et al., 2023) for machine translation. In the multi-modal domain, existing benchmarks such as Flickr8k-EX / CF, Composite, PASCAL-50S, THUMB (Kasai et al., 2022), and Polaris (Wada et al., 2024) target image captioning. While some recent benchmarks (Liu et al., 2024; Li et al., 2024) cover multiple tasks, they mainly adopt multiple-choice settings that diverge from real-world text generation. In contrast, our MMHE benchmark provides a multi-task, multi-criteria human evaluation resource, including sentences generated by several state-of-the-art VLMs across diverse tasks. MMHE enables more nuanced evaluations that capture a variety of criteria and tasks, advancing the field beyond the focus on captioning alone.

6. Conclusion

We introduced HarmonicEval, a novel reference-free evaluation metric for multiple multi-modal tasks. HarmonicEval predicts criterion-wise scores and aggregates them via a statistically principled method to produce an overall score. In addition, we constructed MMHE, the first multi-task multi-criteria human evaluation benchmark, consisting of 18,000 expert judgments. Experimental results demon-

strate that HarmonicEval aligns more closely with human judgments than existing metrics on both MMHE and other commonly used human evaluation benchmarks. Furthermore, our analysis with MMHE reveals that existing metrics tend to prioritize certain criteria while neglecting others.

7. Limitations

This section discusses limitations from six perspectives: theory, modality, criteria, data collection, model, and computational cost.

Evaluation Bias. In HarmonicEval, since score distributions are approximated by the output token probabilities, there is a possibility of unintended bias. Notably, evaluation bias in LLM-based evaluation metrics has been documented in several studies (Zheng et al., 2023; Liu et al., 2023b; Ohi et al., 2024). As such, further research into evaluation bias in VLM-based evaluation metrics is essential for future work.

Modality. This study focused primarily on evaluating text quality in vision-language tasks because most state-of-the-art VLMs output only text. This leaves image quality evaluation underexplored. Given that several recent image generation models, such as DALL-E 3, are integrated into conversational systems using VLMs, the automatic evaluation of both generated image and text quality would be a promising next step toward the development of more user-friendly systems.

Criteria. We carefully selected five general criteria that are considered effective across various multi-modal tasks. These criteria were useful for discussions spanning the four multi-modal tasks in this study. To further expand research to include a greater number of criteria and tasks, analyzing the relationship between task- or domain-specific criteria and these general criteria would also be necessary in future work.

Data collection. As the number of criteria increased, it became difficult even for experts to maintain annotation consistency, leading to greater time requirements for data collection. MMHE extracted one hundred images from each task, which was considered to be a statistically reliable number, and each target text was evaluated by three annotators. While large-scale crowdsourcing was attempted to scale up this benchmark, obtaining human judgments that adhered accurately to each rating scale was challenging because careful explanation of the rating process through direct communication was required. This leaves scaling up the number of tasks and images challenging.

Model. Improving VLMs to generate better text based on the evaluation results remains future work. In particular, achieving high scores across all criteria within learning frameworks such as in-context learning or reinforcement learning would be an intriguing direction for further exploration.

Computational cost. Since HarmonicEval relies on five prompts for criterion-wise scoring, it incurs five times the computational or API cost compared to simpler prompting methods. We consider this a trade-off between achieving more robust and fine-grained metrics and managing computational cost. This approach can also be seen as a form of *test-time scaling* (Snell et al., 2025; Xu et al., 2024), where system performance is improved by increasing computational resources at inference time. One potential direction to mitigate this limitation is to have the model generate scores for all criteria in a single output. However, we did not pursue this approach, as it complicates the computation of score expectations and variances. We leave this for future work.

8. Ethics Statement

Data collection. We created and will publicly release a new benchmark as part of this research. The data collection process was conducted with careful consideration for ethical guidelines. All annotators were informed about the purpose of this benchmark and provided consent before participation. Any personally identifiable information has been removed to ensure privacy protection. The benchmark was reviewed to minimize harmful content, offensive language, or biases that could negatively impact downstream applications. However, some inherent biases might still be present due to the nature of the data sources including natural images.

Reproducibility. All code necessary to reproduce the experimental results will be made publicly available. All experiments have been conducted as deterministically as possible by fixing random seeds and setting the temperature hyperparameters to zero.

9. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 25K03135. These research results were obtained from the commissioned research (No.22501) by National Institute of Information and Communications Technology (NICT) , Japan. This work was supported by the “R&D Hub Aimed at Ensuring Transparency and Reliability of Generative

AI Models" project of the Ministry of Education, Culture, Sports, Science and Technology. This study was carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

10. Bibliographical References

Somak Aditya, Yezhou Yang, Chitta Baral, Yiannis Aloimonos, and Cornelia Fermüller. 2018. Image understanding using vision and reasoning through scene description graph. *Elsevier Computer Vision and Image Understanding (CVIU)*, 173:33–45.

Peter Anderson, Basura Fernando, Mark Johnson, et al. 2016. Spice: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 382–398.

Hiroki Asano, Tomoya Mizumoto, and Kentaro Inui. 2017. Reference-based metrics can be replaced with reference-less metrics in evaluating grammatical error correction systems. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 343–348, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 49250–49267.

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

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. InfoMetIC: An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3171–3185, Toronto, Canada. Association for Computational Linguistics.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural*

Language Processing (EMNLP-IJCNLP), pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, Seattle, United States. Association for Computational Linguistics.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, Online. Association for Computational Linguistics.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024a. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11286–11315, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024b. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746, Bangkok, Thailand. Association for Computational Linguistics.

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

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*, page 216–233, Berlin, Heidelberg. Springer-Verlag.

Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. 2024. Vision language model-based caption evaluation method leveraging visual context extraction. *arXiv preprint arXiv:2402.17969*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3190–3199.

Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood-based mitigation of evaluation bias

in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3237–3245, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! find one mismatch between image and language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 255–265, Vancouver, Canada. Association for Computational Linguistics.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2025. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained summarization evaluation using LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 906–922, Bangkok, Thailand. Association for Computational Linguistics.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. 2025. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):7419–7427.

Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13559–13568.

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

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2024b. Cogvlm: Visual expert for pretrained language models. In *Proceedings of the Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2024. Llava-cot: Let vision language models reason step-by-step.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v(ision) as a generalist evaluator for vision-language tasks.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Haojie Zhuang, Wei Emma Zhang, Leon Xie, Weitong Chen, Jian Yang, and Quan Sheng. 2024. Automatic, meta and human evaluation for multimodal summarization with multimodal output. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7768–7790, Mexico City, Mexico. Association for Computational Linguistics.