VELA: An LLM-Hybrid-as-a-Judge Approach for Evaluating Long Image Captions

Kazuki Matsuda Yuiga Wada Shinnosuke Hirano Seitaro Otsuki Komei Sugiura
Keio University
{k2matsuda0, yuiga, shinhirano, otsu8sei14, komei.sugiura}@keio.jp

Abstract

In this study, we focus on the automatic evaluation of long and detailed image captions generated by multimodal Large Language Models (MLLMs). Most existing automatic evaluation metrics for image captioning are primarily designed for short captions and are not suitable for evaluating long captions. Moreover, recent LLM-as-a-Judge approaches suffer from slow inference due to their reliance on autoregressive inference and early fusion of visual information. To address these limitations, we propose VELA, an automatic evaluation metric for long captions developed within a novel LLM-Hybrid-as-a-Judge framework. Furthermore, we propose LongCap-Arena, a benchmark specifically designed for evaluating metrics for long captions. This benchmark comprises 7,805 images, the corresponding human-provided long reference captions and long candidate captions, and 32,246 human judgments from three distinct perspectives: *Descriptiveness*, *Relevance*, and *Fluency*. We demonstrated that VELA outperformed existing metrics and achieved superhuman performance on LongCap-Arena. Our code and dataset are available at https://vela.kinsta.page/.

1 Introduction

Multimodal Large Language Models (MLLMs) have been widely researched and applied in various social domains, including robotics and healthcare. (Achiam et al., 2023; Team et al., 2023; Liu et al., 2023, 2024a; Lin et al., 2024; Dai et al., 2023; Bai et al., 2024). To effectively develop MLLMs, it is essential to use automatic evaluation metrics that closely align with human judgments. Despite the proficiency of MLLMs in generating long and detailed captions, effective evaluation metrics for assessing this capability have not yet been fully established. Indeed, classic metrics, such as BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), have been shown to exhibit weak correlation with human judgments when evaluating long captions.

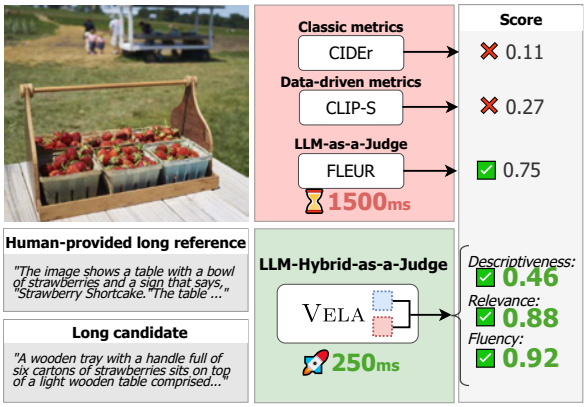

Figure 1: Overview of VELA, which evaluates long image captions from three perspectives: *Descriptiveness*, *Relevance*, and *Fluency*. VELA employs an LLM-Hybrid-as-a-Judge framework, which enables both computational efficiency and high alignment with human judgments.

In this study, we focus on the automatic evaluation of long and detailed image captions generated by MLLMs. Our target task is challenging even for humans, as we will demonstrate in Section 5.2.

Although there have been many attempts to develop evaluation metrics for image captions, they remain inadequate for evaluating long captions. Indeed, recent approaches (Hessel et al., 2021; Sarto et al., 2023, 2024b; Wada et al., 2024; Matsuda et al., 2024) exhibit low correlation with human judgments. Moreover, LLM-as-a-Judge approaches (Chan et al., 2023; Lee et al., 2024; Tong et al., 2025; Yao et al., 2024) are impractical because of their slow inference. In fact, they require over three hours to evaluate generated captions on standard benchmarks (e.g., (Lin et al., 2014; Agrawal et al., 2019)). This is largely attributed to the autoregressive nature of LLM-based inference and the early fusion of visual information.

To address these limitations, we propose VELA, an automatic evaluation metric for long image cap-

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 8680–8696
November 4-9, 2025

tions, developed within a novel LLM-Hybrid-as-a-Judge framework. Fig. 1 presents an overview of VELA with a typical sample. To train and validate the proposed metric, we construct the LongCap-Arena benchmark, which includes images, human-provided long reference captions, long candidate captions, and human judgments.

VELA distinguishes itself from existing metrics in two key aspects: First, VELA adopts a late fusion approach to integrate visual information with a non-autoregressive LLM, unlike existing metrics (e.g. (Lee et al., 2024; Tong et al., 2025)). This late fusion approach avoids increases in the input sequence lengths, enabling inference that is faster than that of early fusion approaches. Second, instead of outputting a single score that represents the overall quality of a candidate, the proposed metric outputs evaluation scores across three distinct perspectives: *Descriptiveness* (*Desc.*), *Relevance* (*Rel.*), and *Fluency* (*Flu.*) This prevents certain evaluation criteria from being ignored, which is a common issue in metrics that output only a single score (Ohi et al., 2024).

The main contributions of this study are summarized as follows:

1. We propose VELA, a supervised metric evaluating long image captions from three distinct perspectives.
2. We introduce an LLM-Hybrid-as-a-Judge framework, which enables computationally efficient and LLM-based evaluations while incorporating images through a Reference-to-Candidate LLM (R2C-LLM) branch and an Image-to-Candidate Alignment (I2C-Align) branch.
3. We construct LongCap-Arena, a benchmark for both training and evaluating metrics on long captions, featuring 32,246 human judgments collected from 1,020 annotators.
4. VELA outperformed existing metrics, including LLM-as-a-Judge approaches, and achieved superhuman performance on the LongCap-Arena benchmark.

2 Related Work

Several survey papers on MLLMs and evaluation for image captioning (Stefanini et al., 2022; Ghandi et al., 2023; Caffagni et al., 2024; Gu et al., 2025) provide comprehensive overviews of standard models and automatic evaluation metrics. In particular, (Gu et al., 2025) provided a broad summary of LLM-as-a-Judge approaches across various text generation tasks, including image captioning.

Image captioning metrics. Standard automatic evaluation metrics for image captioning include BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016). Extensions such as CIDEr-R (Oliveira et al., 2021) and JaSPICE (Wada et al., 2023) have been proposed. Although these classic metrics have been widely used in image captioning, researchers have shown that they weakly correlate with human judgments (Hessel et al., 2021; Sarto et al., 2023, 2024b; Wada et al., 2024; Matsuda et al., 2024).

As a result, data-driven evaluation metrics (Lee et al., 2020, 2021; Hessel et al., 2021; Sarto et al., 2023, 2024b; Wada et al., 2024; Matsuda et al., 2024) have been proposed that leverage pretrained models. However, these metrics primarily target short captions and are not suitable for evaluating long captions. HiFiScore (Yao et al., 2024) is one of the few metrics targeting long captions, which transforms both images and candidate captions into hierarchical parsing graphs and evaluates the candidates based on node-level matching between the corresponding graphs. Although it performs well on short captions (e.g., (Aditya et al., 2015; Hodosh et al., 2013)), converting both the caption and image into hierarchical parsing graphs can lead to information loss when evaluating long captions.

LLM-as-a-Judge approaches. Automatic evaluation metrics based on LLMs or MLLMs, often referred to as LLM-as-a-Judge approaches, have been shown to be successful across a variety of evaluation tasks (Gu et al., 2025). Several LLM-as-a-Judge approaches have also been proposed for the automatic evaluation of image captioning. For example, FLEUR (Lee et al., 2024) uses an MLLM (LLaVA (Liu et al., 2023, 2024a)) and performs early fusion of the visual inputs to enable evaluation with an image input. Similarly, G-VEval (Tong et al., 2025) and HarmonicEval (Ohi et al., 2024) also employ MLLMs to incorporate visual information and provide more interpretable evaluations by scoring captions from multiple perspectives. Despite their advantages, MLLM-based metrics often suffer from slow inference, which results from both the increase in input sequence length caused by early fusion of the visual inputs and the use

of autoregressive inference. Indeed, these metrics require over three hours to evaluate generated captions on standard benchmarks (e.g., COCO (Lin et al., 2014), nocaps (Agrawal et al., 2019)). Such inefficiency leads to practical issues in the use of these metrics during the development of MLLMs.

Datasets and benchmarks. Standard datasets for evaluating image captioning metrics include Composite (Aditya et al., 2015), Flickr8K-Expert, Flickr8K-CF (Hodosh et al., 2013), Polaris (Wada et al., 2024), and Nebula (Matsuda et al., 2024). However, these datasets provide human judgments from a single evaluation perspective only, limiting their ability to capture the diverse quality dimensions of the candidates.

Several recent studies have proposed datasets that include multi-dimensional human judgments (Ohi et al., 2024; Kasai et al., 2022). MMHE (Ohi et al., 2024) provides 4,500 human judgments for 100 images across five evaluation perspectives: *Correctness*, *Completeness*, *Clarity*, *Fluency*, and *Conciseness*. Similarly, THumB (Kasai et al., 2022) includes 2,500 human judgments for 500 images, across two dimensions: *Precision* and *Recall*. However, all these datasets are limited to short captions; Composite, Flickr8K-CF, Polaris, and THumB have average caption lengths of 12.6, 11.4, 9.4, and 10.2 words, respectively. Although the number of datasets for evaluating long captions remains limited, ParaEval (Yao et al., 2024) is a representative example. ParaEval is based on the ImageParagraph dataset (Krause et al., 2017), which contains 4,000 images paired with long references. For each reference, it provides automatically generated negative samples based on four error types: *plausible*, *attribute*, *object*, and *relation*.

3 Problem Statement

Automatic evaluation for long captions. We focus on the automatic evaluation of long and detailed image captions generated by MLLMs. Fig. 2 illustrates an automatic evaluation of long captions. In this task, given an image $\boldsymbol{x}_{\text{img}}$, a long candidate $\boldsymbol{x}_{\text{cand}}$, and N human-provided long references $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$, automatic evaluation metrics assess $\boldsymbol{x}_{\text{cand}}$ in relation to both $\boldsymbol{x}_{\text{img}}$ and $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$ from three perspectives. It is desirable for the metrics to output scores that closely align with human judgments.

These three perspectives are defined as follows:

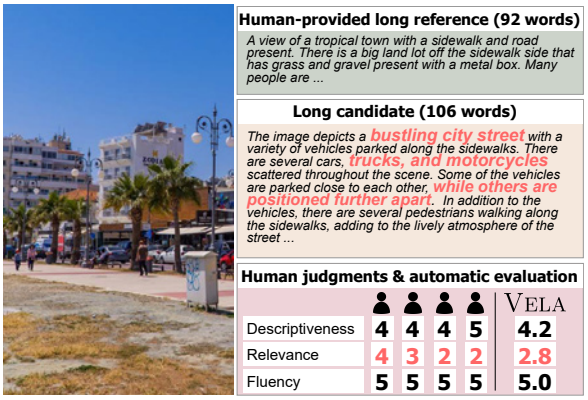

Figure 2: Example of automatic evaluation for long captions. In this task, automatic evaluation metrics assess a candidate based on the given image and human-provided long references across three perspectives: *Descriptiveness*, *Relevance*, and *Fluency*. The evaluation scores should align with human judgments.

- ***Descriptiveness (Desc.)*** evaluates how thoroughly the candidate caption $\boldsymbol{x}_{\text{cand}}$ captures the details of the image $\boldsymbol{x}_{\text{img}}$, including objects, attributes, and relationships.
- ***Relevance (Rel.)*** measures the extent to which $\boldsymbol{x}_{\text{cand}}$ appropriately reflects the content of $\boldsymbol{x}_{\text{img}}$, by identifying errors such as incorrect objects (e.g., "dog" instead of "cat"), attributes (e.g., "blue" instead of "red"), and relationships (e.g., "under" instead of "on top of").
- ***Fluency (Flu.)*** focuses on the grammatical correctness, coherence, and naturalness of $\boldsymbol{x}_{\text{cand}}$, including any grammatical or spelling errors, redundancy, and unnecessary phrases that affect linguistic quality.

We identified these three perspectives based on the conventions of various natural language generation tasks such as image captioning (Lee et al., 2021; Aditya et al., 2015; Kasai et al., 2022; Yao et al., 2024; Liu et al., 2019; Yue et al., 2023), text summarization (Kryscinski et al., 2019; Fabbri et al., 2021; Song et al., 2024), and machine translation (Freitag et al., 2021).

4 Method

We propose VELA, an automatic evaluation metric tailored for evaluating long and detailed image captions. Fig. 3 shows the architecture of VELA. It consists of two main branches: the R2C-LLM branch and I2C-Align branch.

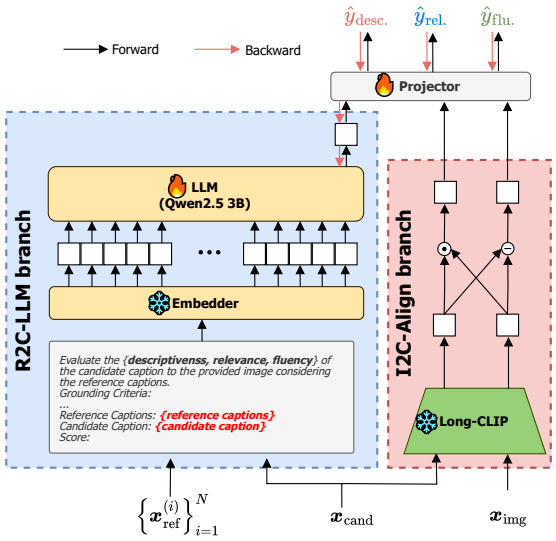

Figure 3: Architecture of VELA. The image, long candidate, and human-provided long references are processed by our metric through two branches: R2C-LLM and I2C-Align. The R2C-LLM branch leverages an LLM to capture the linguistic relationship between the candidate and references, whereas the I2C-Align branch uses Long-CLIP to compute the similarity between the candidate and image.

4.1 R2C-LLM branch.

This branch efficiently assesses the quality of $\boldsymbol{x}_{\text{cand}}$ in relation to the corresponding $\boldsymbol{x}_{\text{ref}}$ by employing a lightweight LLM in a non-autoregressive manner. We adopt evaluation based on LLMs to take advantage of their extensive world knowledge and linguistic capability acquired through pretraining on broad-domain datasets. To address the slow speed of autoregressive inference in LLM-as-a-Judge, we employ a non-autoregressive approach that significantly reduces the inference time. Furthermore, although MLLMs typically perform early fusion of visual information, which results in increased computational costs and slow inference, we opt for a text-only LLM with a late fusion approach to mitigate these issues.

In the R2C-LLM branch, we first prepare a prompt $\boldsymbol{x}_{\text{prompt}}$ using $\boldsymbol{x}_{\text{cand}}$ and $\{\boldsymbol{x}_{\text{ref}}^{(i)}\}_{i=1}^{N}$. Our evaluation prompt, designed based on previous work (Hodosh et al., 2013; Fabbri et al., 2021; Tong et al., 2025; Lee et al., 2021), is provided in Appendix J. We then feed $\boldsymbol{x}_{\text{prompt}}$ into a text-only LLM (Qwen2.5-3B (Yang et al., 2024)), and obtain the last hidden states in a non-autoregressive manner. The sequence of hidden states is denoted as $\{\boldsymbol{h}_i\}_{i=1}^{M}$, where M denotes the sequence length. Similar to previous works using the last hidden states (e.g., (BehnamGhader et al., 2024; Su et al., 2023; Springer et al., 2025)), we compute $\boldsymbol{g}_{\text{r2c}}$, the output of the R2C-LLM branch, as follows:

$$\boldsymbol{g}_{\text{r2c}} = \left[\frac{1}{M}\sum_{i=1}^{M}\boldsymbol{h}_i \;,\; \boldsymbol{h}_M\right].$$

4.2 I2C-Align branch.

This branch evaluates $\boldsymbol{x}_{\text{cand}}$ with respect to $\boldsymbol{x}_{\text{img}}$ using Long-CLIP (Zhang et al., 2024) without relying on MLLMs. As previously mentioned, the early fusion of visual information in MLLM-based metrics results in high computational costs (Chan et al., 2023; Lee et al., 2024; Tong et al., 2025). To avoid these costs, the I2C-Align branch does not employ MLLMs.

The I2C-Align branch uses Long-CLIP to extract $\boldsymbol{h}_{\text{img}}$ and $\boldsymbol{h}_{\text{cand}}$ from $\boldsymbol{x}_{\text{img}}$ and $\boldsymbol{x}_{\text{cand}}$, respectively. Unlike existing metrics based on CLIP (Hessel et al., 2021; Sarto et al., 2023, 2024b,a; Wada et al., 2024; Matsuda et al., 2024), the I2C-Align branch employs Long-CLIP to overcome the 77-token limit of the original CLIP model, which is insufficient for processing long captions that typically exceed 100 words.

The output of I2C-Align ($\boldsymbol{g}_{\text{i2c}}$) is then computed as follows:

$$\boldsymbol{g}_{\text{i2c}} = \left[\left|\boldsymbol{h}_{\text{img}} - \boldsymbol{h}_{\text{cand}}\right| \;,\; \boldsymbol{h}_{\text{img}} \odot \boldsymbol{h}_{\text{cand}}\right],$$

where $\left|\boldsymbol{h}_{\text{img}} - \boldsymbol{h}_{\text{cand}}\right|$ and $\boldsymbol{h}_{\text{img}} \odot \boldsymbol{h}_{\text{cand}}$ denote the absolute element-wise difference and Hadamard product between $\boldsymbol{h}_{\text{img}}$ and $\boldsymbol{h}_{\text{cand}}$, respectively. These operations have been shown to be effective in automatic evaluation across various text generation tasks, such as machine translation and image captioning (Shimanaka et al., 2018; Rei et al., 2020; Wada et al., 2024; Matsuda et al., 2024).

The final scores $\hat{\boldsymbol{y}} \in \mathbb{R}^3$ are computed as follows:

$$\hat{\boldsymbol{y}} = (\hat{y}_{\text{desc}}, \hat{y}_{\text{rel}}, \hat{y}_{\text{flu}}) = \sigma\left(\boldsymbol{W}\left[\boldsymbol{g}_{\text{r2c}}, \boldsymbol{g}_{\text{i2c}}\right] + \boldsymbol{b}\right),$$

where σ denotes the sigmoid function, and $\boldsymbol{W}$ and $\boldsymbol{b}$ are trainable parameters. Here, $\hat{y}_{\text{desc}}$, $\hat{y}_{\text{rel}}$, and $\hat{y}_{\text{flu}}$ denote the predicted scores for *Desc.*, *Rel.*, and *Flu.*, respectively. We employed the mean squared error as our loss function.

5 Experiments

5.1 Experimental Setup

LongCap-Arena benchmark. We constructed LongCap-Arena, a benchmark specifically designed to evaluate metrics for long image captions.

To the best of our knowledge, few datasets specifically focus on evaluating metrics for long captions (Yao et al., 2024). Existing long-caption datasets for metric evaluation (e.g., ParaEval(Yao et al., 2024)) do not contain human-provided references annotated based on the semantic structures of images (Urbanek et al., 2024; Krause et al., 2017; Pont-Tuset et al., 2020). Moreover, candidates in ParaEval are limited to either human-provided references or negative examples generated through simple word replacements, limiting diversity in candidate quality.

To address these limitations, we constructed LongCap-Arena, a benchmark that enables comprehensive and authorized evaluation by providing candidate captions with diverse quality and human-provided long references derived from the DCI dataset (Urbanek et al., 2024). This benchmark comprises images, long candidate captions, human-provided long reference captions, and human judgments obtained by assessing long candidate captions from three perspectives. Unlike existing datasets with human judgments (Aditya et al., 2015; Hodosh et al., 2013; Wada et al., 2024), LongCap-Arena contains captions with over 100 words on average. Moreover, LongCap-Arena provides human judgments from multiple perspectives, in contrast to most existing image captioning datasets, which assess only the overall appropriateness of the candidates.

To construct LongCap-Arena, we used images and long reference captions from the DCI dataset (Urbanek et al., 2024). The DCI dataset includes comprehensive, high-quality, human-provided captions that describe nearly every element within an image. These detailed captions closely reflect the visual content, making them a reliable basis for evaluating the quality of long candidate captions. However, the DCI dataset only provides pairs of images and references. Because it lacks candidates and the corresponding human judgments, we cannot directly use this dataset for evaluating metrics.

Therefore, we collected long candidate captions generated by ten different models and gathered human judgments for each image–candidate pair. For each image, we generated long candidates using ten representative MLLMs and image captioning models. We employed the same prompts for each respective MLLM to generate these candidates.

Subsequently, each pair of candidates and images was independently assessed by human annotators from three perspectives: *Desc.*, *Rel.*, and *Flu.*. Following previous studies (Achiam et al., 2023; Dai et al., 2023; Chen et al., 2024b; Liu et al., 2024b,a; Gong et al., 2023; Bai et al., 2024; Chen et al., 2024a; Li et al., 2023; Wang et al., 2022), the annotators assessed candidates on a five-point scale based on detailed guidelines. To support the evaluation of *Desc.*, we utilized SAM (Kirillov et al., 2023) to generate object masks corresponding to image regions. These object masks served as visual cues to assist the annotators in determining the necessary level of detail in their evaluations.

To align with the DCI's split of the validation and test sets, we divided the VELA test set into two subsets: TestA and TestB. TestA comprises all images from the DCI dataset's validation set, whereas TestB includes all images from the DCI dataset's test set. Further details on the benchmark and its construction process are provided in Appendix B.

5.2 Quantitative Results

Table 1 presents a quantitative comparison with baseline metrics on the TestA and TestB sets. Following previous research on image captioning metrics (Sarto et al., 2023, 2024b; Tong et al., 2025; Sarto et al., 2024a; Zeng et al., 2024), we adopted Kendall's τ_b and τ_c to evaluate the metrics. Due to space constraints, Table 1 only displays the results for Kendall's τ_c. Results for τ_b are provided in Appendix D. Table 1 also compares the inference time per sample for each evaluation metric.

We adopted BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), CLIP-S (Hessel et al., 2021), RefCLIP-S (Hessel et al., 2021), PAC-S (Sarto et al., 2023), RefPAC-S (Sarto et al., 2023), Polos (Wada et al., 2024), DENEB (Matsuda et al., 2024), FLEUR (Lee et al., 2024), Ref-FLEUR (Lee et al., 2024), PAC-S++ (Sarto et al., 2024b), RefPAC-S++ (Sarto et al., 2024b), and G-VEval (Tong et al., 2025) as baselines because they are standard, representative metrics in the field of automatic evaluation for image captioning.

Metrics such as CLIP-S, PAC-S, and PAC-S++, which are based on CLIP (Radford et al., 2021), have a maximum input text length limited to 77 tokens by CLIP. Therefore, for a fair comparison, we followed the approach in (Yao et al., 2024) and employed modified versions of CLIP-S, PAC-S, and PAC-S++ (called CLIP-S$_{avg}$, PAC-S$_{avg}$, and PAC-S++$_{avg}$, respectively) as baselines alongside their original implementations. Specifically, these modified versions computed the cosine similarity

	Metrics	TestA $[\tau_c] \uparrow$			TestB $[\tau_c] \uparrow$			Inference time [ms] $\downarrow$
		Desc.	*Rel.*	*Flu.*	*Desc.*	*Rel.*	*Flu.*	
Image captioning metrics	BLEU	28.6	2.4	25.5	32.0	-10.1	-3.5	0.46
	CIDEr	-7.0	6.7	4.4	4.0	-3.4	1.9	1.3
	CLIP-S	24.5	18.6	25.5	27.3	22.5	24.5	26
	CLIP-S$_{avg}$	-8.6	11.5	3.2	12.8	27.5	28.4	200
	RefCLIP-S	13.4	7.3	9.5	21.2	10.3	10.9	33
	PAC-S	24.8	14.7	23.6	27.6	25.7	23.0	48
	PAC-S$_{avg}$	-7.4	14.6	6.2	6.6	29.2	28.4	360
	RefPAC-S	22.6	19.1	24.9	40.7	29.2	27.9	52
	Polos	28.5	18.1	30.6	41.1	22.4	20.0	33
	DENEB	10.3	18.4	22.2	31.3	35.7	<u>32.6</u>	47
	PAC-S++	29.7	21.4	34.2	28.1	21.9	21.1	36
	PAC-S++$_{avg}$	-7.2	19.4	6.0	14.1	32.4	30.3	270
	RefPAC-S++	25.4	23.3	28.9	40.3	22.2	24.2	40
LLM-as-a-Judge	FLEUR	17.3	2.6	0.5	12.6	10.6	-3.1	1300
	RefFLEUR	21.3	10.3	7.2	28.1	12.3	17.5	1400
	G-VEval	28.3	22.5	18.2	38.1	22.2	19.2	1800
	GPT4o w/o references	<u>54.1±1.0</u>	<u>36.8±6.3</u>	20.9±1.0	43.6±2.0	<u>37.3±3.4</u>	25.2±1.0	1900
	GPT4o w/ references	47.0±1.1	26.2±2.2	<u>35.4±2.9</u>	<u>46.9±2.6</u>	30.4±2.3	25.1±4.3	2000
	VELA (Ours)	**56.4±1.3**	**40.0±1.1**	**57.4±1.3**	**54.0±0.4**	**52.3±1.1**	**39.0±2.3**	260
	Human performance	56.1	46.6	24.5	48.9	52.6	24.4	—

Table 1: Quantitative comparison with baseline metrics. **Bold** font indicates the best results and <u>underlined</u> font indicates the second best results. ViT-L/14 is used as the backbone for metrics that rely on CLIP (Radford et al., 2021).

between each sentence in the paragraph and the image, then outputted the final score by calculating the average of these scores.

Furthermore, we evaluated the performance of GPT-4o under both reference-free and reference-based settings. For GPT-4o w/o references and GPT-4o w/ references, we utilized a modified version of the FLEUR (Lee et al., 2024) prompt, specifically tailored to assess the three aspects *Desc.*, *Rel.*, and *Flu.* For both VELA and GPT-4o (with and without reference), we report the mean and standard deviation over five runs.

Correlation with human judgments. Table 1 demonstrates that our proposed metric achieved scores of 56.4, 40.0, and 57.4 for *Desc.*, *Rel.*, and *Flu.* on the TestA set, and 54.0, 52.3, and 39.0 on the TestB set, respectively.

VELA outperformed both the reference-free and reference-based versions of GPT-4o on the TestA set by 2.3 points in *Desc.*, 3.2 points in *Rel.*, and 22.0 points in *Flu.* Similarly, on the TestB set, VELA achieved improvements of 5.3 points in *Desc.*, 1.7 points in *Rel.*, and 15.6 points in *Flu.*, compared with all baseline metrics.

The differences in τ_c between the proposed metric and each baseline metric were statistically significant ($p < 0.05$) for *Rel.* and *Flu.* on the TestA set, and for *Desc.*, *Rel.*, and *Flu.* on the TestB set.

Inference time. Table 1 also shows the inference times per sample on our LongCap-Arena benchmark, evaluated using a GeForce RTX 3090 GPU and an Intel Core i9-10900KF CPU. LLM-free metrics such as CLIP-S$_{avg}$, PAC-S$_{avg}$, and PAC-S++$_{avg}$ demonstrated inference times ranging from approximately 1 ms to 400 ms. Moreover, existing LLM-based metrics such as FLEUR, RefFLEUR, G-VEval, and GPT-4o exhibited significantly longer inference times of 1280 ms, 1392 ms, 1812 ms, and 1905 ms, respectively—all values exceeding 1000 ms. In contrast, VELA achieved an inference time of only 258 ms, which is approximately five times faster than the LLM-based metrics. These measurements include the time for both tokenization and CUDA kernel launches for a fair comparison.

5.3 Qualitative Analysis

In the example on the left, although $\boldsymbol{x}_{\text{cand}}$ captured the primary visible elements of the image, it lacked the detailed description found in $\boldsymbol{x}_{\text{ref}}$. Therefore, the human annotators assigned it a *Desc.* score of 0.55. Although GPT-4o w/o references overestimated the *Desc.* score at 0.8, VELA evaluated it more appropriately with a score of 0.43, which aligns closely with the human judgment. Note that $\boldsymbol{x}_{\text{cand}}$ contained repetitions (e.g., "The stadium is used for football matches"), which led the human

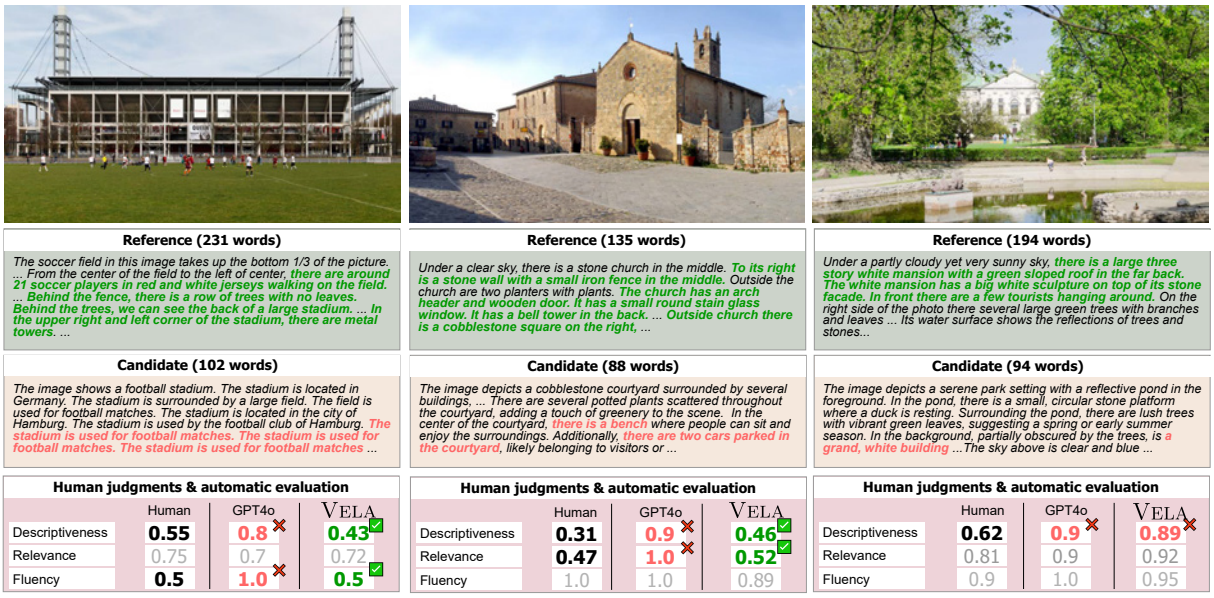

Figure 4: Qualitative results on LongCap-Arena. The left and middle subfigures illustrate successful cases, while the right subfigure shows a failure case. Each subfigure consists of $\boldsymbol{x}_{\text{img}}$, $\boldsymbol{x}_{\text{ref}}$ ("Reference"), $\boldsymbol{x}_{\text{cand}}$ ("Candidate"), and human judgments $\boldsymbol{y}$ along with automatic evaluation scores $\hat{\boldsymbol{y}}$ ("Human judgments & automatic evaluation"). Values in green and red indicate scores that are closely aligned and misaligned with human judgments, respectively.

annotators to assign a *Flu.* score of 0.5. In contrast, GPT-4o overestimated the *Flu.* score, assigning a score of 1.0, whereas VELA evaluated it correctly, assigning a score of 0.5.

The middle subfigure presents another successful example for VELA. Here, $\boldsymbol{x}_{\text{cand}}$ primarily described the dominant elements of the image but failed to capture the details described in $\boldsymbol{x}_{\text{ref}}$; therefore, the human annotators assigned it a *Desc.* score of 0.31. However, GPT-4o overestimated the DESC., assigning a score of 0.9, whereas VELA evaluated it more appropriately at 0.46. In the sample, $\boldsymbol{x}_{\text{cand}}$ included hallucinated objects (e.g., "a bench" and "two cars parked in the courtyard"), leading human annotators to assign a *Rel.* score of 0.47. GPT-4o failed to evaluate this correctly, assigning a score of 0.9. In contrast, VELA evaluated it more appropriately, assigning a score of 0.46.

In the right-hand subfigure of Fig. 4, $\boldsymbol{x}_{\text{ref}}$ provided a detailed description of the white building at the center of the image, as indicated by the green text. Although the building occupied a relatively small region, this level of detail was likely motivated by its central placement and the absence of other semantically important objects. In contrast, $\boldsymbol{x}_{\text{cand}}$ described it only as "a grand white building," which lacked the level of detail provided in $\boldsymbol{x}_{\text{ref}}$. Given that the human judgment for *Desc.* was 0.62, it is desirable for the automatic evaluation metrics not to overestimate the quality of this candidate. However, both the proposed metric and GPT-4o assigned inappropriately high scores for *Desc.*, at 0.89 and 0.9, respectively.

To identify the cause, we examined the output of the R2C-LLM branch, without fusing it with the output of the I2C-Align branch. Although the fused output yielded a higher score of 0.89, the R2C-LLM branch outputted a score of 0.75, which was closer to the human judgment. This result suggests that the I2C-Align branch may have contributed to the discrepancy, possibly because it failed to recognize the white building as a key object.

5.4 Ablation Studies

Table 2 shows the quantitative results of the ablation studies. We conducted three ablation studies to investigate the contribution of each module in our proposed metric.

LLM-Hybrid Ablation. We investigated the contribution of the LLM-Hybrid-as-a-Judge framework by excluding each branch. As shown in Table 2, a comparison between Metric (i) and Metric (viii) indicates that excluding the R2C-LLM branch led to decreases of 10.5, 28.1, and 45.8 points on TestA and 29.4, 23.5, and 34.1 points on TestB for *Desc.*, *Rel.*, and *Flu.*, respectively. In contrast, Metric (ii), which excludes the I2C-Align branch, also showed performance decreases of 6.2, 4.2, and 1.1 points on TestA and 5.5, 2.8, and 0.5 points on TestB for

Metric	R2C-LLM backbone	LLM-Hybrid	I2C-Align backbone	TestA [τ_c] ↑			TestB [τ_c] ↑		
				Desc.	***Rel.***	***Flu.***	***Desc.***	***Rel.***	***Flu.***
(i)	—	—	Long-CLIP ViT-L/14	45.9	11.9	11.6	24.6	28.8	4.9
(ii)	Qwen2.5-3B	—	—	50.2	35.8	56.3	48.5	49.5	38.8
(iii)	Qwen2.5-3B	✓	CLIP ViT-L/14	54.9	37.2	51.5	50.0	46.7	34.5
(iv)	Qwen2.5-3B	✓	PAC-S CLIP ViT-L/14	54.4	37.7	53.1	51.9	48.0	32.1
(v)	Qwen2.5-3B	✓	Long-CLIP ViT-L/14	56.0	39.4	57.2	51.6	49.1	34.1
(vi)	Llama3.2-3B	✓	Long-CLIP ViT-L/14	54.5	36.2	51.8	51.0	49.8	**41.3**
(vii)	Phi-3.5-Mini	✓	Long-CLIP ViT-L/14	51.0	30.7	54.8	44.7	48.1	32.2
(viii)	Qwen2.5-3B	✓	Long-CLIP ViT-L/14	**56.4±1.3**	**40.0±1.1**	**57.4±1.3**	**54.0±0.4**	**52.3±1.1**	39.0±2.3

Table 2: Results of ablation studies on the effect of incorporating the LLM-Hybrid-as-a-Judge framework and using different R2C-LLM and I2C-Align backbones. These results demonstrated that integrating the Long-CLIP ViT-L/14 backbone and LLM-Hybrid-as-a-Judge framework significantly contributed to the performance.

Desc., *Rel.*, and *Flu.*, respectively. These results demonstrate that both branches contribute to the performance improvement.

I2C-Align Backbone Ablation. We analyzed the impact of different backbones in the I2C-Align branch by replacing the Long-CLIP ViT-L/14 backbone with alternative models. Table 2 shows that Metric (viii) outperformed Metric (iii) with the CLIP ViT-L/14 backbone, Metric (iv) with the PAC-S CLIP ViT-L/14 backbone, and Metric (v) with the Long-CLIP ViT-B/16 backbone.

Specifically, Metric (viii) achieved improvements of 2.1, 6.3, and 6.9 points in *Desc.*, *Rel.*, and *Flu.* on the TestB set compared with Metric (iv) using the PAC-S CLIP ViT-L/14 backbone, respectively. These results demonstrate that the Long-CLIP ViT-L/14 backbone played a crucial role in enhancing performance.

R2C-LLM Backbone Ablation. We investigated the effect of different R2C-LLM backbones by replacing the Qwen2.5-3B backbone with Llama3.2-3B and Phi-3.5 Mini. We selected these models because they were lightweight yet high-performing, with model sizes comparable to Qwen2.5-3B. Table 2 shows that Metric (viii) outperformed Metric (v) with the Llama3.2-3B backbone and Metric (vii) with the Phi-3.5 Mini backbone. Specifically, on the TestA set, Metric (viii) achieved improvements of 1.9, 3.8, and 5.6 points in *Desc.*, *Rel.*, and *Flu.*, respectively, compared with the results of Metric (vi) using the Llama3.2-3B backbone, respectively. These results demonstrate that the Qwen2.5-3B backbone played a crucial role in enhancing performance.

5.5 Comparison with Human Performance

We conducted a subject experiment to evaluate human performance on the TestA and TestB sets. Six participants participated in the experiment and were divided into two groups of three. Each group was assigned to evaluate the three aspects of long captions on either the TestA set or TestB set.

We calculated Kendall's τ_c for each evaluator's judgments against the ground truth in our dataset, and then computed the average Kendall's τ across all evaluators to measure the human performance. As shown in Table 1, the human performance results for *Desc.*, *Rel.*, and *Flu.* were 55.1, 41.0, and 28.1 on the TestA set, and 48.7, 50.6, and 23.4 for the TestB set, respectively.

Table 1 shows that the proposed metric outperformed human evaluation in both *Desc.* and *Flu.* Specifically, the proposed metric outperformed human evaluation by 0.3 and 5.1 points in *Desc.*, and by 32.9 and 14.6 points in *Flu.*, on the TestA and TestB sets respectively. These results indicate that the proposed metric achieved performance comparable to human evaluation in assessing the *Desc.* of candidates. Moreover, the large margin in *Flu.* suggests that the proposed metric could potentially replace human evaluation in assessing naturalness and grammatical correctness.

By contrast, Table 1 shows that in *Rel.*, the proposed method underperformed human performance by 6.6 points on the TestA set and 0.3 points on the TestB set. As discussed in Appendix E, this performance gap could be attributed to insufficient grounding in the I2C-Align branch and the suboptimal integration of outputs from the R2C-LLM and I2C-Align branches. A possible solution to this is to integrate visual information while leveraging a pretrained language model, similar to the gated xattn-dense layer in Flamingo (Alayrac et al., 2022).

6 Conclusion

In this study, we focused on the automatic evaluation of long and detailed image captions generated

by MLLMs. The contributions of this study are as follows: (i) We proposed VELA, a supervised metric evaluating long image captions from three distinct perspectives. (ii) We introduced the LLM-Hybrid-as-a-Judge framework, which enables computationally efficient and LLM-based evaluations while incorporating images through the R2C-LLM and I2C-Align branches. (iii) We constructed LongCap-Arena, a benchmark designed for both training and evaluating metrics on long captions, featuring 32,246 human judgments collected from 1,020 annotators. (iv) VELA outperformed existing metrics and achieved superhuman performance on the LongCap-Arena benchmark.

7 Limitations

Although our metric has clearly been shown to provide a high correlation with human judgments, it is not without its limitations. The primary limitation is the occurrence of errors stemming from the lack of sufficient detail or accuracy in the references. Moreover, the metric tends to erroneously overlook semantically important objects in the image, especially when they occupy relatively small regions. These limitations could be attributed to insufficient grounding in the I2C-Align branch and the suboptimal integration of outputs from the R2C-LLM and I2C-Align branches. Another important limitation is that the R2C-LLM branch requires access to last hidden states, which prevents the direct use of closed-source models. For further error analysis, see Appendix E

Acknowledgments

This work was supported by a grant from Apple Inc. Any views, opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and should not be interpreted as reflecting the views, policies, or position, either expressed or implied, of Apple Inc. This work was also partially supported by JSPS KAKENHI Grant Number 23K28168 and JST Moonshot.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. *arXiv preprint arXiv:1511.03292*.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. nocaps: Novel Object Captioning at Scale. In *ICCV*, pages 8948–8957.

Jean Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A Visual Language Model for Few-shot Learning. In *NeurIPS*, volume 35, pages 23716–23736.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *ECCV*, pages 382–398.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2024. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. In *ICLR*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *ACL*, pages 65–72.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large Language Models Are Secretly Powerful Text Encoders. In *COLM*.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The Revolution of Multimodal Large Language Models: A Survey. In *ACL*, pages 13590–13618.

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *EMNLP*, pages 13638–13646.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. In *ECCV*, pages 370–387.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *CVPR*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat 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 *NeurIPS*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *TACL*, 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. *TACL*, 9:1460–1474.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep Learning Approaches on Image Captioning: A Review. *ACM Computing Surveys*, 56(3).

Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. 2023. MultiModal-GPT: A Vision and Language Model for Dialogue with Humans. *arXiv preprint arXiv:2305.04790*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. *arXiv preprint arXiv:2411.15594*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *EMNLP*, pages 7514–7528.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *JAIR*, 47:853–899.

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 *NAACL*, pages 3464–3478.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *ICCV*, pages 3992–4003.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei. 2017. A Hierarchical Approach for Generating Descriptive Image Paragraphs. In *CVPR*, pages 3337–3345.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural Text Summarization: A Critical Evaluation. In *EMNLP*, pages 540–551.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *ACL*, pages 220–226.

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 *Evaluation and Comparison of NLP Systems*, pages 34–39.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In *ACL*, pages 3732–3746.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, pages 19730–19742.

Chin Lin. 2004. ROUGE: A Package For Automatic Evaluation Of Summaries. In *ACL*, pages 74–81.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: On Pre-training for Visual Language Models. In *CVPR)*, pages 26679–26689.

Tsung Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollar. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved Baselines with Visual Instruction Tuning. In *CVPR*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Jae Lee. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *NeurIPS*, pages 34892–34916.

Lixin Liu, Jiajun Tang, Xiaojun Wan, and Zongming Guo. 2019. Generating Diverse and Descriptive Image Captions Using Visual Paraphrases. In *ICCV*, pages 4239–4248.

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. 2024. DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning. In *ACCV*, pages 3570–3586.

Masanari Ohi, Masahiro Kaneko, Naoaki Okazaki, and Nakamasa Inoue. 2024. HarmonicEval: Multimodal, Multi-task, Multi-criteria Automatic Evaluation Using a Vision Language Model. *arXiv preprint arXiv:2412.14613*.

Gabriel Oliveira, Esther Colombini, and Sandra Avila. 2021. CIDEr-R: Robust Consensus-based Image Description Evaluation. In *W-NUT*, pages 351–360.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *ACL*, pages 311–318.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting Vision and Language with Localized Narratives. In *ECCV*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, pages 8748–8763.

Ricardo Rei, Craig Stewart, Ana Farinha, and Alon Lavie. 2020. COMET: A Neural Framework for MT Evaluation. In *EMNLP*, pages 2685–2702.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *CVPR*, pages 6914–6924.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024a. BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues. In *ECCV*, page 70–87.

Sara Sarto, Moratelli Nicholas, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024b. Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. *arXiv preprint arXiv:2410.07336*.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2018. RUSE: Regressor Using Sentence Embeddings for Automatic Machine Translation Evaluation. In *WMT*, pages 751–758.

Hwanjun Song, Hang Su, Igor Shalyminov, Jason Cai, and Saab Mansour. 2024. FineSurE: Fine-grained Summarization Evaluation using LLMs. In *ACL*, pages 906–922.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2025. Repetition Improves Language Model Embeddings. In *ICLR*.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2022. From Show to Tell: A Survey on Deep Learning-based Image Captioning. *TPAMI*, 45(1):539–559.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One Embedder, Any Task: Instruction-Finetuned Text Embeddings. In *ACL*, pages 1102–1121.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, and Amelia Glaese. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

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. In *AAAI*, pages 7419–7427.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions. In *CVPR*, pages 26700–26709.

Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, pages 4566–4575.

Yuiga Wada, Kanda Kaneda, and Komei Sugiura. 2023. JaSPICE: Automatic Evaluation Metric Using Predicate-Argument Structures for Image Captioning Models. In *CoNLL*, pages 424–435.

Yuiga Wada, Kaneda Kanta, Saito Daichi, and Komei Sugiura. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *CVPR*, pages 13559–13568.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. GIT: A Generative Image-to-text Transformer for Vision and Language. *TMLR*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Ziwei Yao, Ruiping Wang, and Xilin Chen. 2024. HiFi-Score: Fine-Grained Image Description Evaluation with Hierarchical Parsing Graphs. In *ECCV*, pages 441–458.

Zihao Yue, Anwen Hu, Liang Zhang, and Qin Jin. 2023. Learning Descriptive Image Captioning via Semipermeable Maximum Likelihood Estimation. In *NeurIPS*, pages 3462–3479.

Zequn Zeng, Jianqiao Sun, Hao Zhang, Tiansheng Wen, Yudi Su, Yan Xie, Zhengjue Wang, and Bo Chen. 2024. HICEScore: A Hierarchical Metric for Image Captioning Evaluation. In *ACM*, page 866–875.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-CLIP: Unlocking the Long-Text Capability of CLIP. In *ECCV*, pages 310–325.

A Additional Related Work

CLIP-S (Hessel et al., 2021) and PAC-S (Sarto et al., 2023) use CLIP (Radford et al., 2021) to compute the cosine similarity between the vector representations of the image and candidate. In contrast, Polos (Wada et al., 2024) and DENEB (Matsuda et al., 2024) employ supervised learning based on human judgments, achieving high correlation with human judgments. CLAIR (Chan et al., 2023) is an LLM-based metric that uses GPT-3.5 to evaluate candidates with respect to human-provided references. However, it does not incorporate visual information, which limits its applicability to image-grounded tasks.

B Construction of LongCap-Arena

In this study, we constructed LongCap-Arena, a new benchmark designed for evaluating metrics for long captions. LongCap-Arena contains 32,246 human judgments and 7,805 images, each paired with human-provided long reference captions, and long candidate captions. The candidate captions were generated by ten representative MLLMs and image captioning models based on the DCI dataset images. These models include GPT-4o (Achiam et al., 2023), InstructBLIP (Dai et al., 2023), InternVL (Chen et al., 2024b), LLaVA-NeXT (Liu et al., 2024b), LLaVA-1.5 (Liu et al., 2024a), MultimodalGPT (Gong et al., 2023), Qwen-VL-Chat (Bai et al., 2024), ShareGPT4V (Chen et al., 2024a), BLIP2 (Li et al., 2023), and GIT (Wang et al., 2022).

The candidate captions included 7,805 captions with a vocabulary size of 21,611 words, a total word count of 570,600, and an average length of 101.2 words. Moreover, the reference captions consisted of 7,805 captions with a vocabulary size of 20,988 words, a total word count of 738,848, and an average length of 131.4 words. Moreover, all captions are in English.

This dataset is built on the DCI dataset (Urbanek et al., 2024), which provides images and long reference captions annotated by humans. When creating a new dataset from existing images in the DCI dataset, it is crucial to carefully address the issue of potential data leakage in MLLMs, which could arise if the training set of the DCI dataset is also used to train the MLLMs. Therefore, we constructed the training and validation sets of the VELA dataset using the training set of the DCI dataset.

The training, validation, TestA, and TestB sets contain 11,971, 1,309, 294, and 324 samples, respectively. The training set was used for training the metric, the validation set for hyperparameter tuning, and the test set for evaluating the metric's performance.

Human judgments were provided on a five-point scale and subsequently normalized to the range $[0, 1]$. Each candidate caption was evaluated by a minimum of three distinct annotators. The final score for each caption was determined by calculating the average of these individual judgments. Fig. 5 illustrates the annotation interface used for evaluating *Desc*. The annotation was conducted via a public crowdsourcing platform, where we recruited annotators from a general population on the internet without restricting demographic or geographic background. We recruited annotators and provided payment that was adequate based on the participants' country of residence, and obtained consent via the task instructions, which clearly stated that the collected data would be used for research purposes. To ensure data reliability, we excluded responses exhibiting suspicious behavior, such as excessively short response times or repeated identical scores.

C Implementation Details

Epoch	10
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
Learning rate	1.0×10^{-4}
Batch size	4

Table 3: Settings of the proposed metric.

Table 3 shows the training settings of the proposed metric. Our metric had approximately 3.68 million trainable parameters. Our metric was trained on a system equipped with an NVIDIA GeForce RTX 3090 GPU with 24GB memory and an Intel Core i9-12900K CPU with 64GB RAM. The training process was completed within approximately three hours. We employed early stopping during training using Kendall's τ_c. Specifically, τ_c was computed on the validation set after each epoch. Training was terminated when no improvement in τ_c was observed on the validation set for a single epoch. Subsequently, the metric's performance was assessed on the test set.

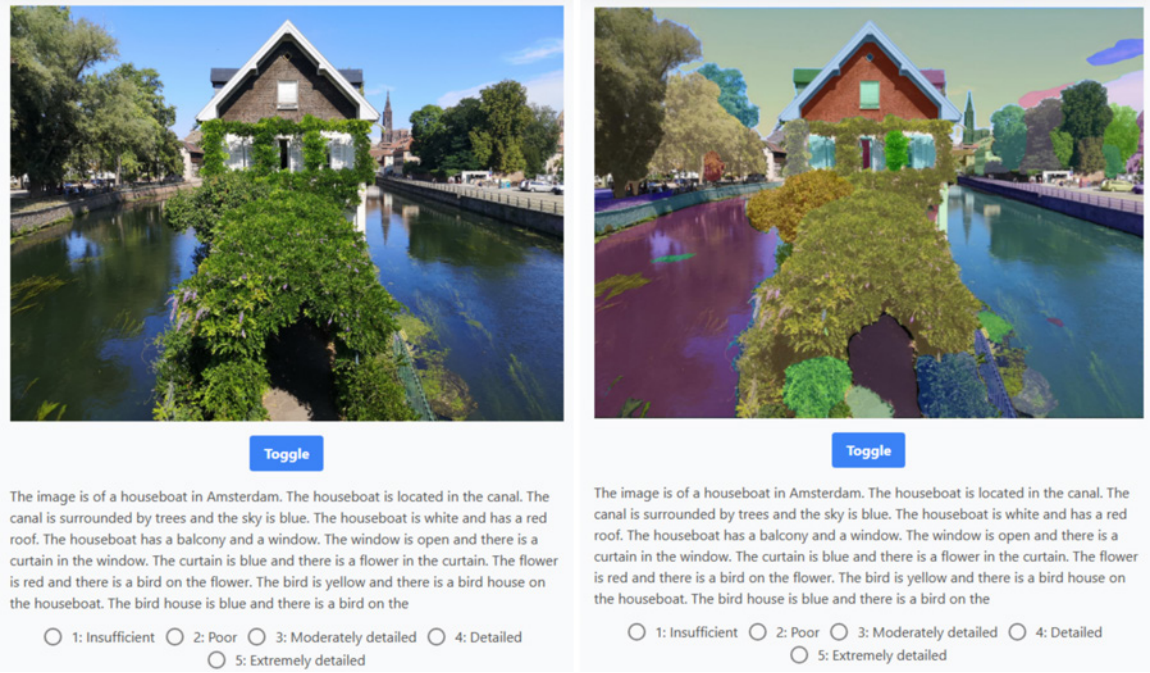

Figure 5: Annotation interface for *Desc*. The left subfigure shows the normal image, and the right subfigure presents its segmented version generated using SAM (Kirillov et al., 2023). These object masks were shown to annotators as visual cues, helping them determine the level of detail required for their evaluation.

Metrics	**TestA** $[\tau_b] \uparrow$			**TestB** $[\tau_b] \uparrow$		
	Desc.	***Rel.***	***Flu.***	***Desc.***	***Rel.***	***Flu.***
GPT4o w/o references	56.1±1.3	38.2±6.2	35.2±1.0	44.6±1.8	38.8±4.1	**39.3±3.0**
GPT4o w/ references	47.4±1.1	27.3±2.3	37.0±3.2	47.2±2.6	31.7±2.3	27.1±5.3
VELA (Ours)	**56.8±1.1**	**41.0±1.2**	**55.2±0.9**	**52.2±1.8**	**51.9±1.2**	38.7±2.1
Human performance	54.4	47.5	32.1	46.6	51.9	31.1

Table 4: Quantitative results (Kendall's τ_b) on the LongCap-Arena benchmark. VELA outperforms GPT-4o (w/ and w/o references) across *Desc.*, *Rel.*, and *Flu.*, and notably surpasses human performance in *Desc.* and *Flu.*

D Quantitative Results for Kendall's τ_b

Table 4 presents additional results using Kendall's τ_b, confirming that VELA achieved superior performance compared with GPT-4o (with and without references) on both the TestA and TestB sets.

E Error Analysis

To investigate the limitations of the proposed metric, we analyzed samples on which the proposed metric did not perform as expected. We defined failure cases for each evaluation perspective as samples where the corresponding output satisfied the following condition:

$$|y_{\text{per}} - \hat{y}_{\text{per}}| \geq \theta, \quad \text{per} \in \{\text{desc}, \text{rel}, \text{flu}\}$$

In this study, we fixed θ at a value of 0.25, as this value corresponds to the difference between two adjacent points on a normalized five-point scale. Under this condition, we identified 9, 11, and 5 failure cases in *Desc.*, *Rel.*, and *Flu.*, respectively, from the combined TestA and TestB sets, which comprise 206 samples in total.

Table 5 shows the categorization of failure modes, which are grouped into the following categories:

- Insufficient detail or accuracy in references:
  This category encompasses failure modes where the references lack sufficient image-related details, leading to evaluation discrepancies in *Desc.*
- Redundant candidates:
  This category refers to failure modes where

Failure Mode Category	#Errors
Insufficient detail or accuracy in references	12
Redundant candidates	2
Over-reliance on references	3
Named entities in candidates	1
Fluency issues	3
Short candidates	2
Others	2
Total	25

Table 5: Categorization of failure modes.

the proposed metric has assigned inappropriate scores to candidates with redundant information unrelated to the image.

- Incorrect or missing information in references:
  This category pertains to failure modes where the references contain incorrect or missing information, leading to evaluation discrepancies in *Rel.*
- Over-reliance on references:
  This category refers to failure modes where the proposed metric has prioritized the references over the image, leading to evaluation scores that differ from human judgments.
- Named entities in candidates:
  This category refers to failure modes where candidates include named entities whose correctness cannot be determined solely based on the image, leading to evaluation discrepancies.
- Fluency issues:
  This category encompasses failure modes where the proposed metric has assigned scores that do not align with human judgments to candidates containing unnatural elements (e.g., unnatural phrasing, grammatical or spelling errors, redundancy, or extraneous characters).
- Short candidates:
  This category refers to failure modes where the proposed metric has assigned inappropriate scores to short candidates because the training data predominantly consisted of long captions.
- Others:
  This category encompasses various errors that do not fall into the aforementioned categories.

We independently categorized failure modes in each evaluation perspective into the above categories. Table 5 shows that the primary cause of errors was the lack of sufficient detail or accuracy in the references. These errors likely arise because the proposed metric does not effectively handle the outputs from the I2C-Align branch, which do not rely on references, and fails to adequately integrate the outputs of the R2C-LLM and I2C-Align branches. In future work, we plan to extend our metric by introducing a mechanism that integrates visual information while leveraging a pretrained language model, similar to the gated xattn-dense layer in Flamingo (Alayrac et al., 2022).

F Fusion of Visual and Textual Features

To investigate potential limitations in expressivity, we conducted experiments on the comparison of Transformer-based and MLP-based fusion between visual and textual features. Table 6 shows the results of the variant modified to employ Transformer-based fusion. The results demonstrate that the Transformer-based fusion achieved performance comparable to the MLP-based fusion overall, with slightly lower scores in *Desc*. Given these results and the computational cost, we adopted MLP-based fusion in the final model.

G VELA in Reference-free Setting

VELA can be used in a reference-free setting by simply removing the reference input from the R2C-LLM prompt, without any modification to the architecture. Table 7 shows the results of evaluating VELA in the reference-free setting. Although the absence of human-annotated references led to a decrease in performance compared to the reference-based VELA, reference-free VELA outperformed GPT-4o in both settings. Specifically, reference-free VELA demonstrates an improvement over reference-free GPT-4o by +0.3, -0.7, and +35.5 points on TestA, and by +2.8, +19.2, and +8.2 points on TestB, in *Desc.*, *Rel.*, and *Flu.*, respectively. These results demonstrate that VELA can generalize well to real-world scenarios where reference captions are unavailable.

H Zero-shot Evaluation on Short Caption Benchmarks

To evaluate potential overfitting to the DCI dataset (Urbanek et al., 2024), we additionally conducted experiments in a zero-shot setting on two standard benchmarks for short caption evaluation: Composite (Aditya et al., 2015) and Flickr8k-

Metrics	TestA $[\tau_c] \uparrow$			TestB $[\tau_c] \uparrow$		
	Desc.	***Rel.***	***Flu.***	***Desc.***	***Rel.***	***Flu.***
VELA w/ Transformer	53.3	39.5	57.6	52.0	52.3	37.0
VELA w/ MLP (Ours)	**56.4±1.3**	**40.0±1.1**	**57.4±1.3**	**54.0±0.4**	**52.3±1.1**	**39.0±2.3**

Table 6: Comparison of Transformer-based and MLP-based fusion between visual and textual features. While the Transformer-based fusion achieved competitive results, the MLP-based fusion demonstrated better performance in *Desc.*

Metrics	TestA $[\tau_c] \uparrow$			TestB $[\tau_c] \uparrow$		
	Desc.	***Rel.***	***Flu.***	***Desc.***	***Rel.***	***Flu.***
GPT-4o w/o references	54.1±1.0	36.8±6.3	20.9±1.0	43.6±2.0	37.3±3.4	25.2±1.0
GPT-4o w/ references	47.0±1.1	26.2±2.2	35.4±2.9	46.9±2.6	30.4±2.3	25.1±4.3
VELA w/o references	54.3	36.1	56.4	46.4	44.4	33.4
VELA w/ references (Ours)	**56.4±1.3**	**40.0±1.1**	**57.4±1.3**	**54.0±0.4**	**52.3±1.1**	**39.0±2.3**

Table 7: Quantitative results of VELA in a reference-free setting. Reference-free VELA outperformed GPT-4o in both reference-free and reference-based settings.

Expert (Hodosh et al., 2013). In these experiments, we modified VELA to output a single score instead of three perspective scores, following the evaluation protocol of these benchmarks. Table 8 shows that VELA achieved competitive performance with existing state-of-the-art metrics (Vedantam et al., 2015; Hessel et al., 2021; Sarto et al., 2023; Wada et al., 2024; Chan et al., 2023; Lee et al., 2024; Tong et al., 2025) on both benchmarks. These results indicate that VELA was not overfitted to the DCI dataset and remains effective for short caption evaluation.

I Scoring Criteria for Annotation

Descriptiveness:
Annotators were instructed to evaluate how detailed the caption was in describing the image content, focusing on objects, relationships, and attributes. They used both the *"normal image"* and *"segmented image"* (with color-coded objects) to assess the level of detail.
The scoring criteria were as follows:

5: Extremely detailed — The caption comprehensively describes all observed objects and relationships, including spatial relationships and contextual details.

4: Detailed — The caption describes most objects and relationships, with only minor omissions.

3: Moderately detailed — The caption mentions key objects but lacks detail in relationships or other attributes.

2: Poor — The caption includes descriptions of a few objects but omits significant details and relationships.

1: Insufficient — The caption provides minimal or no description of the image content.

Relevance:
Relevance was evaluated based on the correctness of objects, attributes, and relationships mentioned in the captions.
Proper nouns and associated specific details (e.g., *"Mt. Fuji"*) were excluded from the evaluation.
The scoring criteria were as follows:

5: Fully relevant — The caption appropriately describes the image content without errors.

4: Mostly relevant — Minor inaccuracies are present but the overall caption is almost correct.

3: Partially relevant — Significant inaccuracies exist, yet some parts of the caption remain correct.

2: Barely relevant — Numerous inaccuracies significantly distort the relevance of the caption.

1: Not relevant — The caption is fundamentally unrelated to the image content.

Metric	Composite [τ_c] ↑	Flickr8k-Expert [τ_c] ↑
CIDEr (Vedantam et al., 2015)	37.7	43.9
CLIP-S (Hessel et al., 2021)	53.8	51.2
RefCLIP-S (Hessel et al., 2021)	55.4	53.0
PAC-S (Sarto et al., 2023)	55.7	54.3
RefPAC-S (Sarto et al., 2023)	57.3	55.9
Polos (Wada et al., 2024)	57.6	56.4
CLAIR (Chan et al., 2023)	–	48.8
FLEUR (Lee et al., 2024)	**63.5**	53.0
G-VEval (Tong et al., 2025)	–	**59.7**
VELA (Ours)	61.3	56.2

Table 8: Results of zero-shot evaluation on the short caption benchmarks, Composite and Flickr8k-Expert. VELA achieved comparable performance with existing state-of-the-art metrics.

Fluency:
Annotators were directed to evaluate the naturalness and grammatical correctness of captions, independent of their accuracy.
Markdown syntax (e.g., ###, -) was not considered an error.
The scoring criteria were as follows:

5: Extremely fluent — No errors or minimal errors (no more than one).
4: Fluent — Minor errors present but the caption is generally natural and comprehensible.
3: Moderately fluent — Noticeable errors are present but the text remains understandable.
2: Lacking fluency — Numerous errors make the caption difficult to read.
1: Not fluent — Frequent errors render the caption incomprehensible.

J Prompts in VELA

This section provides the full prompts used in the R2C-LLM branch of VELA for *Desc.*, *Rel.*, and *Flu.*

J.1 Descriptiveness

`System`
Evaluate the descriptiveness of the candidate caption based on the reference captions and the provided image. Focus only on how detailed the caption is, regardless of relevance. Refer to the following criteria:
- 5: Extremely detailed - Captures all objects, relationships, and attributes in the image with precise and complete descriptions, including spatial relationships and overall context.
- 4: Detailed - Captures most objects and relationships but lacks some elements.
- 3: Partially detailed - Mentions key objects but misses spatial relationships or additional details.
- 2: Insufficient detail - Mentions only a few objects correctly, with many elements missing.
- 1: Very poor detail - Mentions almost no objects and fails to represent the image content.
Only give a number from 1 to 5 with no text.

`User`
Reference Captions: {{Reference}}
Candidate Caption: {{Candidate}}

`Assistant`
Score:

J.2 Relevance

`System`
Evaluate the relevance of the candidate caption to the provided image considering the reference captions. Focus solely on how well the caption aligns with the image content, ignoring fluency or descriptiveness. Refer to the following criteria:
- 5: Fully relevant - Accurately describes the image content with no errors.
- 4: Mostly relevant - Contains minor errors but is generally aligned with the image content.
- 3: Partially relevant - Includes significant errors but some parts relate to the image.
- 2: Barely relevant - Contains many errors and deviates significantly from the image content.

- 1: Not relevant - Contains numerous errors and fundamentally mismatches the image.
Only give a number from 1 to 5 with no text.
User

Reference Captions: {{Reference}}
Candidate Caption: {{Candidate}}

Assistant
Score:

J.3 Fluency

System
Evaluate the fluency of the candidate caption, focusing solely on its grammatical correctness, naturalness, and readability. Ignore the content's relevance or descriptiveness. Refer to the following criteria:
- 5: Very fluent - No errors or only one minor error; reads naturally as proper English sentences.
- 4: Fluent - Contains some errors but is overall natural and easy to understand.
- 3: Partially fluent - Noticeable errors but still comprehensible.
- 2: Lacking fluency - Many errors that make it hard to read.
- 1: Not fluent - Excessive errors that make it incomprehensible.
Only give a number from 1 to 5 with no text.

User
Reference Captions: {{Reference}}
Candidate Caption: {{Candidate}}

Assistant
Score:

K Additional Details for ARR Checklist

Discuss The License For Artifacts VELA and LongCap-Arena are released under the BSD 3-Clause Clear License.

The licenses of the models and datasets used in this study are summarized below:

DCI dataset (Urbanek et al., 2024): CC BY-NC 4.0
InstructBLIP (Dai et al., 2023): Research (non-commercial)
InternVL (Chen et al., 2024b): MIT license
LLaVA-NeXT (Liu et al., 2024b): Apache 2.0 license
LLaVA-1.5 (Liu et al., 2024a): Apache 2.0 license
Multimodal-GPT (Gong et al., 2023): Apache 2.0 license
Qwen-VL-Chat (Bai et al., 2024): Tongyi Qianwen License
ShareGPT4V (Chen et al., 2024a): Apache 2.0 license
BLIP2 (Li et al., 2023): BSD 3-Clause license
GIT (Wang et al., 2022): MIT license
Long-CLIP (Zhang et al., 2024): Apache 2.0 license
Qwen2.5 (Yang et al., 2024): Apache 2.0 license

Artifact Use Consistent With Intended Use All existing artifacts used in this study were utilized in a manner consistent with their intended use. For the artifacts we created, we define their intended use as general academic and research use, which is compatible with the original access conditions of the datasets and models employed in this study.

Data Contains Personally Identifying Info Or Offensive Content The collected data contain no personally identifiable or offensive content. All data used in this study are publicly available. We also confirmed that the source websites, repositories, and publications include no statements indicating concerns about personal information.