DISCODE: Distribution-Aware Score Decoder for Robust Automatic Evaluation of Image Captioning

Nakamasa Inoue[1]*, Kanoko Goto[1]*, Masanari Oi[1], Martyna Gruszka[1], Mahiro Ukai[1], Takumi Hirose[1], Yusuke Sekikawa[2]

[1]Institute of Science Tokyo
[2]Denso IT Laboratory

Abstract

Large vision-language models (LVLMs) have shown impressive performance across a broad range of multimodal tasks. However, robust image caption evaluation using LVLMs remains challenging, particularly under domain-shift scenarios. To address this issue, we introduce the **Distribution-Aware Score Decoder (DISCODE)**, a novel finetuning-free method that generates robust evaluation scores better aligned with human judgments across diverse domains. The core idea behind DISCODE lies in its test-time adaptive evaluation approach, which introduces the Adaptive Test-Time (ATT) loss, leveraging a Gaussian prior distribution to improve robustness in evaluation score estimation. This loss is efficiently minimized at test time using an analytical solution that we derive. Furthermore, we introduce the **Multi-domain Caption Evaluation (MCEval) benchmark**, a new image captioning evaluation benchmark covering six distinct domains, designed to assess the robustness of evaluation metrics. In our experiments, we demonstrate that DISCODE achieves state-of-the-art performance as a reference-free evaluation metric across MCEval and four representative existing benchmarks.

1 Introduction

Developing automatic evaluation metrics that closely correlate with human judgments is essential for advancing toward more human-centric artificial intelligence. For image captioning tasks, significant efforts have been devoted to designing automatic evaluation metrics, beginning with traditional methods such as BLEU (Papineni et al. 2002) and CIDEr (Vedantam, Zitnick, and Parikh 2015). Nevertheless, accurate evaluation remains challenging due to the inherent variability and subjectivity of natural language descriptions.

Recently, large vision-language models (LVLMs) have demonstrated substantial improvements in image-text alignment tasks (Liu et al. 2023a; Wang et al. 2024a; Chen et al. 2024b; Lu et al. 2024). To perform accurate numerical evaluation using LVLMs, state-of-the-art methods such as FLEUR (Lee, Park, and Kang 2024) and G-VEval (Tong et al. 2025) leverage the *score smoothing* technique, which generates real-valued scores by estimating the score distribution based on the output token probability distribution assigned by LVLMs. However, robustly aligning generated scores with human judgments remains difficult, especially in domain-diverse scenarios.

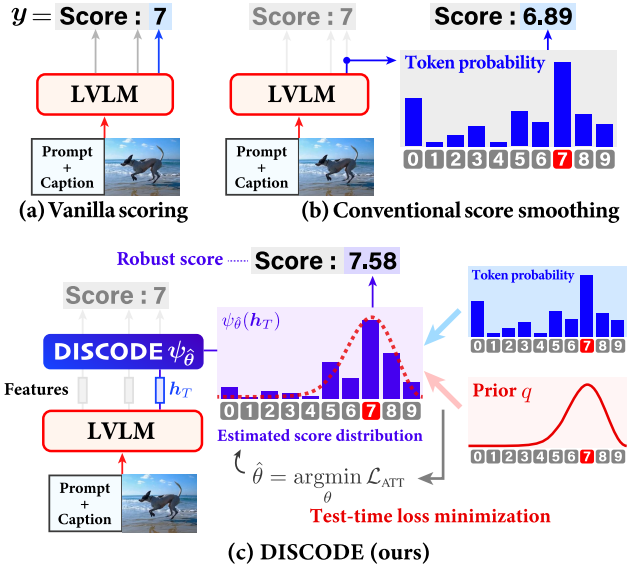

Figure 1: Scoring with DISCODE. (a) Vanilla scoring: LVLM generates raw scores. (b) Score smoothing: Expected score is computed from the token probability distribution. (c) DISCODE (ours): Score distribution is robustly estimated from the decoder feature $\boldsymbol{h}_T$ by minimizing the ATT loss $\mathcal{L}_{\mathrm{ATT}}$, which leverages a Gaussian prior q at test time.

We hypothesize that this difficulty stems from the discrepancy between the token probability distribution and the human evaluation score distribution, particularly in terms of unimodality. Due to the central limit theorem, human evaluation scores naturally tend to follow a Gaussian distribution. In contrast, token probability distributions typically do not exhibit Gaussian behavior and instead show certain biases, such as symbolic bias, where specific tokens are disproportionately frequent (as illustrated in Figure 4). This discrepancy becomes more pronounced in certain visual domains such as paintings and abstract sketches because these domains often involve subjective interpretations and greater

*These authors contributed equally.

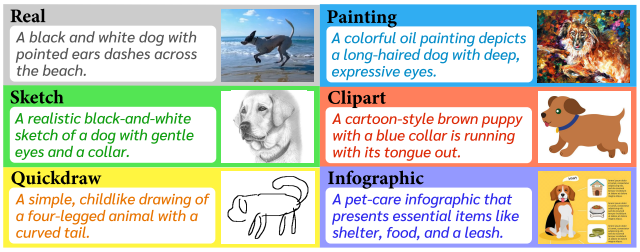

Figure 2: MCEval benchmark. We provide a human evaluation dataset covering six visual domains for assessing the robustness of evaluation metrics.

semantic ambiguity. This motivates us to propose a novel method employing a Gaussian prior distribution to enhance the robustness of evaluation scores, along with a new benchmark dataset encompassing diverse visual domains.

Specifically, this paper makes two major contributions. First, we propose the Distribution-Aware Score Decoder (**DISCODE**), a novel test-time adaptive decoder for LVLMs that generates robust scores by leveraging a Gaussian prior distribution. Second, we introduce the Multi-domain Caption Evaluation (**MCEval**) benchmark, a new image captioning evaluation benchmark spanning six visual domains. In our experiments, we demonstrate that DISCODE achieves state-of-the-art performance on MCEval as well as four representative benchmarks: Flickr8k-Expert (Hodosh, Young, and Hockenmaier 2013), Flickr8k-CF (Hodosh, Young, and Hockenmaier 2013), Composite (Aditya et al. 2015), and Pascal-50S (Vedantam, Zitnick, and Parikh 2015). Our contributions are summarized as follows.

1) Technical Contribution. We propose DISCODE, a novel decoder for LVLM-based image captioning evaluation. As shown in Figure 1, DISCODE minimizes the Adaptive Test-Time (ATT) loss, which measures discrepancy from a Gaussian prior distribution, enabling LVLMs to function as a robust evaluation metric. Furthermore, we derive a closed-form analytical solution to the loss minimization problem, leading to an efficient implementation.

2) Dataset Contribution. We introduce MCEval, a new dataset for benchmarking the robustness and generalizability of evaluation metrics. Our dataset consists of 18,000 image-text pairs spanning six visual domains, as shown in Figure 2, along with human evaluation ground-truth labels.

2 Related Work

Image caption evaluation metrics can be divided into two categories: reference-based and reference-free metrics.

Reference-based metrics. This approach quantifies caption quality by comparing candidate captions with human-written references. Classical metrics such as BLEU (Papineni et al. 2002), ROUGE (Lin 2004), and METEOR (Banerjee and Lavie 2005) rely on n-gram overlap. CIDEr (Vedantam, Zitnick, and Parikh 2015) incorporates TF–IDF weighting to emphasize consensus, and SPICE (Anderson et al. 2016) captures semantic structure by matching scene graphs. With advances in pretrained language models, caption similarity has become measurable in representational space rather than surface form, leading to embedding-based metrics such as BERTScore (Zhang et al. 2020; Yi, Deng, and Hu 2020), BARTScore (Yuan, Neubig, and Liu 2021), TIGEr (Jiang et al. 2019) and ViLBERTScore (Lee et al. 2020). Finetuning-based metrics have further enhanced alignment with human judgments; examples include FAIEr (Wang et al. 2021), Polos (Wada et al. 2024), CAMScore (Cui et al. 2025), DENEB (Matsuda, Wada, and Sugiura 2024)and MLF (Gomes, Zerva, and Martins 2025).

Reference-free metrics. To evaluate image captions without relying on human-written references, reference-free metrics leverage pre-trained vision-language models. CLIPScore (Hessel et al. 2021) is a representative metric that measures the alignment between images and texts using CLIP embeddings (Radford et al. 2021). PAC-S (Sarto et al. 2023, 2024b) introduced positive-augmented contrastive learning, leveraging image generators to more precisely align image and text embeddings. HiFi-Score (Yao, Wang, and Chen 2024) introduced graph-based matching. Recent studies have demonstrated that LVLMs can serve as effective evaluators. To generate accurate real-valued scores, FLEUR (Lee, Park, and Kang 2024) introduced improved score smoothing for LLaVA-1.5 (Liu et al. 2023a). G-VEval (Tong et al. 2025) designed chain-of-thought prompts (Wei et al. 2022) for caption evaluation using GPT-4o. These sophisticated approaches rely on the score smoothing technique (Liu et al. 2023b), which estimates the score distribution based on the output token probability distribution. However, token probability distributions often exhibit unintended biases, such as symbolic bias, causing them to become non-unimodal. DISCODE enhances evaluation robustness by addressing this issue through minimization of the ATT loss, a novel test-time loss formulated with a Gaussian prior distribution.

3 Proposed Method

We introduce DISCODE, a novel test-time adaptive decoder that improves the robustness of LVLM-based image captioning evaluation. By minimizing the ATT loss, which measures divergence between the token probability distribution and a Gaussian prior distribution, DISCODE adaptively generates robust scores at test time. To the best of our knowledge, we are the first to propose a test-time adaptive approach for LVLM-based image captioning evaluation.

3.1 Overview

Problem Setting. The goal of image captioning evaluation is to assign real-valued scores to image-caption pairs that better align with human judgments. We assume the availability of a pre-trained LVLM that, given an appropriate prompt, produces raw evaluation scores s_{raw} in $S = \{0, 1, \cdots, 9\}$. We denote by p_{LVLM} the output token probability distribution over S, extracted at the token index T corresponding to s_{raw}. For example, if the LVLM output is "Score: 7", then T indicates the index of the digit 7, and $p_{\text{LVLM}}(7)$ represents the model's confidence in this score. Thus, the raw score s_{raw} is the digit $s \in S$ with the highest probability $p_{\text{LVLM}}(s)$.

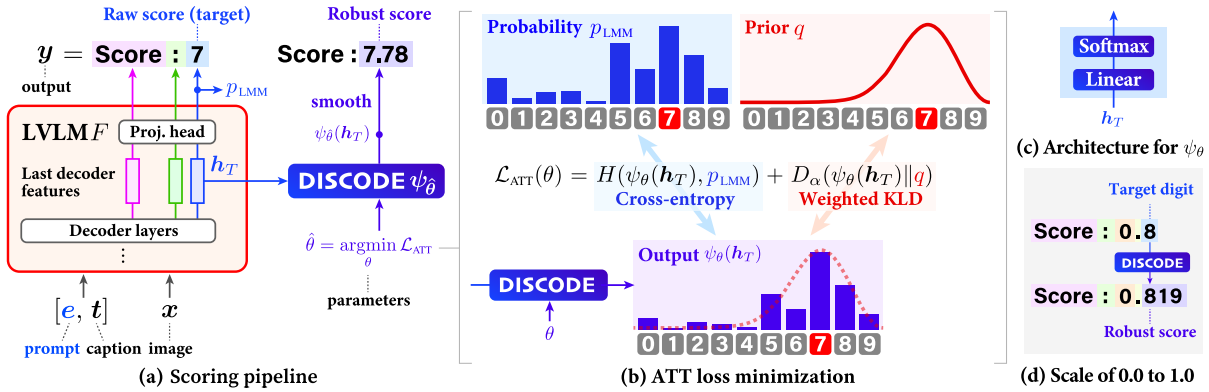

Figure 3: Overview of DISCODE. (a) Scoring pipeline where DISCODE $\psi_{\hat{\theta}}$ is applied to the decoder feature $\boldsymbol{h}_T$ to generate robust scores. (b) ATT loss to minimize cross entropy and weighted Kullback-Leibler divergence (KLD). (c) Architecture for ψ_θ, which consists of a linear layer and a softmax function. (d) Target digit for evaluation on a scale of 0.0 to 1.0.

Scoring Pipeline. Figure 3 (a) shows the scoring pipeline of DISCODE, which consists of three steps. First, given an input image and a candidate caption, the LVLM is prompted to generate a raw score $s_{\text{raw}} \in S$. In this step, we extract the latent feature $\boldsymbol{h}_T \in \mathbb{R}^d$ from the last decoder layer of the LVLM along with the token probability distribution p_{LVLM}. Second, a probability mass distribution $p : S \to [0, 1]$ is estimated, which we refer to as the *score distribution*. This is a core step of the proposed pipeline, where DISCODE generates p by minimizing the ATT loss. Third, the final evaluation score s is computed as the expected value, *i.e.*, $s = \mathbb{E}_{x\sim p(x)}[x]$, thus yielding a real-valued score.

DISCODE. DISCODE ψ_θ is a learnable decoder head to generate the score distribution p from the latent decoder feature $\boldsymbol{h}_T$ as $p = \psi_{\hat{\theta}}(\boldsymbol{h}_T)$. The parameter θ is determined by minimizing the ATT loss:

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \mathcal{L}_{\text{ATT}}(\theta; \boldsymbol{h}_T). \tag{1}$$

This minimization problem is solved independently at test time for each image-caption pair, thereby enabling test-time adaptation across diverse visual domains.

3.2 Adaptive Test-Time Loss

The ATT loss is defined over two probability distributions: a prior distribution q and the token probability distribution p_{LVLM}. As shown in Figure 3(b), minimizing the ATT loss optimizes the balance between these two distributions, resulting in a robust estimation of the score distribution.

Loss Definition. The ATT loss $\mathcal{L}_{\text{ATT}}$ consists of two terms, the cross-entropy term and the divergence term:

$$\mathcal{L}_{\text{ATT}}(\theta; \boldsymbol{h}_T) = \underbrace{H(\psi_\theta(\boldsymbol{h}_T), p_{\text{LVLM}})}_{\text{cross-entropy}} + \underbrace{D_\alpha(\psi_\theta(\boldsymbol{h}_T)\|q)}_{\text{divergence}}, \tag{2}$$

where H is the cross entropy and D_α is the divergence measure. As minimizing the cross-entropy term reduces the discrepancy between the estimated score distribution $\psi_\theta(\boldsymbol{h}_T)$ and p_{LVLM}, the divergence term can be understood as a regularization term to improve the robustness.

Prior Distribution q. Due to the central limit theorem, human evaluation scores naturally tend to follow a Gaussian distribution. To reflect this, we use a Gaussian prior distribution $q(x) \propto \exp\left(-(x - s_{\text{raw}})^2/2\right)$, where $s_{\text{raw}} \in S$ is the raw score generated by the LVLM.

Divergence Term D_α. In numerical evaluations, the highest and lowest scores are often easier to assess than intermediate ones. Consequently, when LVLMs predict scores near the minimum or maximum values, we can rely more on the raw scores and mitigate symbolic bias strongly by assigning greater weight to the unimodal prior. To account for this, we introduce the weighted Kullback-Leibler (KL) divergence for the divergence term, which adaptively determines the dependency on the prior with a parameter $\alpha \in [0, 1]$. Specifically, we define the divergence term by

$$D_\alpha(p\|q) = (1 - \alpha)H(p, q) - \alpha H(p, p), \tag{3}$$

where $H(p, q) = -\sum_{x\in S} p(x) \log q(x)$ is the cross-entropy. The parameter α is adaptively determined based on the raw score using a Gaussian distribution as follows:

$$\alpha = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(s_{\text{raw}} - \mu)^2}{2\sigma^2}\right) \tag{4}$$

where $\mu = |S|^{-1} \sum_{x\in S} x$ is the mean over the candidate digits and $\sigma^2 = 0.1$ is a variance. Note that D_α is equivalent to the vanilla Kullback-Leibler divergence when $\alpha = 0.5$.

Architecture for ψ_θ. We employ a simple yet effective architecture for DISCODE ψ_θ, which consists of a linear layer and a softmax function:

$$\psi_\theta(\boldsymbol{h}) = \text{softmax}(W^\top \boldsymbol{h} + \boldsymbol{b}) \tag{5}$$

where $\theta = \{W, \boldsymbol{b}\}$ is a set of parameters, $W \in \mathbb{R}^{d\times|S|}$ is a weight matrix, $b \in \mathbb{R}^{|S|}$ is a bias vector.

3.3 Analytical Solution

Numerically solving the loss minimization problem in Eq. (1) is computationally expensive. This limitation can be theoretically addressed by deriving an analytical solution. Specifically, the analytical solution of the minimization problem exists for the loss defined in Eq. (2) and the

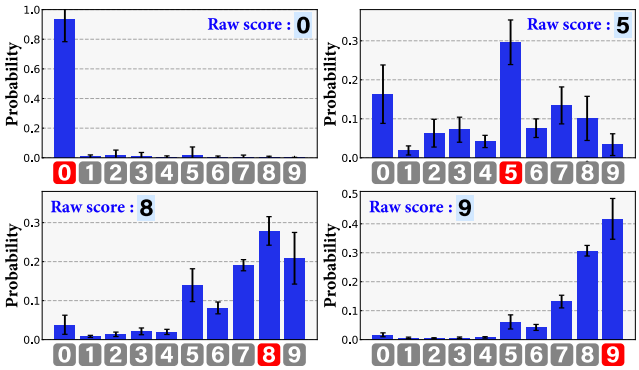

Figure 4: Observed output token probability distributions for four target digits (0, 5, 8 and 9) with LLaVA-Next.

architecture defined in Eq. (5) under the assumption that the LVLM has a linear projection head that predicts token probabilities as $p_{\text{LVLM}} = \text{softmax}(V^\top \boldsymbol{h}_T + \boldsymbol{c})$. The solution is given by $\hat{\theta} = \{\hat{W}, \hat{\boldsymbol{b}}\}$ with

$$\hat{W} = \frac{1}{\alpha} V, \quad \hat{\boldsymbol{b}} = \frac{1-\alpha}{\alpha} \log \boldsymbol{q} + \frac{1}{\alpha} \boldsymbol{c}, \tag{6}$$

where $\boldsymbol{q} \in \mathbb{R}^{10}$ is the vector representation of the prior distribution q. The log function is applied element-wise. The proof is provided in Appendix A.

3.4 Implementation Details

Figure 5 shows the default instruction prompt used in our experiments. This prompt is designed in the FLEUR framework (Lee, Park, and Kang 2024) to produce raw scores on a scale from 0.0 to 1.0. For this prompt, the target is the first decimal place of the raw score. After computing the smoothed score $\hat{s}$, the value below the decimal point in the raw score is replaced with $0.1 \times \hat{s}$ as shown in Figure 3 (d). We build DISCODE on top of ten open-source LVLMs: LLaVA-Next-8B, -13B, -34B, -72B (Li et al. 2024), InternVL-2.5-8B, -78B (Chen et al. 2024a), Qwen2-VL-Instruct-7B, -72B (Wang et al. 2024a), CogVLM2-Chat (Hong et al. 2024) and MiniCPM-V-2.6 (Yao et al. 2024). LLaVA-Next-72B is used as a default LVLM. When using divergences other than KLD for an ablation study, we minimized the loss using the Adam optimizer for ten iterations, with an initial learning rate set to 10^{-3}, because an analytical solution is not available.

3.5 Discussion: Why is DISCODE effective?

Previous studies (Liu et al. 2023b; Lee, Park, and Kang 2024; Tong et al. 2025) assumed that the score distribution p is identical to the output token probability distribution p_{LVLM} assigned by the LVLM, *i.e.*, $p = p_{\text{LVLM}}$. However, p_{LVLM} is not necessarily aligned with the human judgment score distribution, particularly with respect to unimodality. When a sufficiently large number of human evaluators provide scores, the resulting score distribution typically follows a unimodal distribution.[1] In contrast, the token probability distribution may be influenced by unintended bias such as

Instruction prompt

```
Your task is to evaluate and rate the caption on a
scale of 0.0 to 1.0 based on the grading criteria.
Grading criteria:
0.0: Caption does not describe the image at all.
1.0: Caption accurately and clearly describes the image.
Caption: {CAPTION}
Score:
```

Figure 5: Instruction prompt.

Dataset	Domain	Size	Avg length
Flickr8k-Expert	Real	5,664	11.91
Flickr8k-CF	Real	47,830	11.35
Composite	Real	13,146	11.58
Pascal-50S	Real	8,000	8.75
MCEval	**6 domains**	12,000	13.61

Table 1: Comparison with representative image captioning evaluation datasets. Size indicates the number of candidate captions to be evaluated.

symbolic bias. In fact, the probability of the digit 0 is often overestimated as shown in Figure 4, making the distribution non-unimodal. DISCODE is effective because it addresses this issue by robustly estimating the score distribution through minimization of the ATT loss using a Gaussian prior distribution.

4 MCEval Benchmark

This section presents MCEval, a novel human evaluation dataset for benchmarking the robustness and generalizability of evaluation metrics. MCEval consists of 6,000 images, each with two candidate captions and one reference caption, totaling 18,000 image-caption pairs across six domains.

Dataset Construction. We constructed MCEval using images from DomainNet (Peng et al. 2019) and InfographicVQA (Mathew et al. 2022), covering six visual domains: *real*, *painting*, *sketch*, *quickdraw*, *clipart*, and *infograph*. Our annotation process involved three steps. First, we randomly sampled 1,000 images for each domain and generated their candidate captions using four proprietary LVLMs: GPT-4o-mini, GPT-4o, Gemini 2.0 Flash, and Claude 3.5 Sonnet. Open-source LVLMs were excluded because DISCODE depends on them. Second, we randomly selected one candidate caption per image, and human annotators revised it to form a reference caption. Third, for each image, three annotators compared two other candidate captions, selecting the better caption based on relevance, descriptiveness, correctness, and fluency. If consensus among these annotators was not reached, the image was discarded and replaced with another randomly sampled image from the same domain. Annotation tasks were completed by 81 crowdworkers via MTurk and Upwork platforms.

[1] More precisely, human evaluations may become polarized into two opposing extremes, resulting in a convex but not unimodal distribution. However, such cases rarely occur within the scope of quality evaluation.

	Metric	LVLM	FF	Real	Painting	Sketch	Quickdraw	Clipart	Infograph	Mean
Reference-based	BLEU-4 (Papineni et al. 2002)		–	59.3	58.1	61.5	55.7	54.9	53.2	57.1
	ROUGE (Lin 2004)		–	57.9	56.9	59.5	54.5	54.4	50.6	55.6
	METEOR (Banerjee and Lavie 2005)		–	67.8	64.8	67.5	61.8	60.8	59.4	63.7
	CIDEr (Vedantam et al. 2015)		–	66.7	64.5	68.7	62.8	64.5	60.2	64.6
	BERT-S (Zhang et al. 2020)		✓	68.5	73.8	74.3	72.6	67.0	58.6	69.1
	Polos (Wada et al. 2024)			81.3	75.0	77.6	76.8	74.5	69.0	75.7
Reference-free	CLIP-S (Hessel et al. 2021)		✓	79.2	78.0	78.3	75.4	73.9	66.7	75.3
	PAC-S (Sarto et al. 2023)			80.7	71.1	69.7	67.5	66.8	58.7	69.1
	FLEUR (Lee et al. 2024)	✓	✓	84.7	83.6	80.4	45.6	79.9	86.0	76.7
	G-VEval (GPT-4o) (Tong et al. 2025)	✓	✓	86.0	80.2	81.2	76.9	80.6	81.0	81.0
	FLEUR†	✓	✓	86.9	84.3	83.1	76.3	82.0	82.3	82.5
	DISCODE (Ours)	✓	✓	**87.8**	**85.2**	**83.9**	**78.5**	**83.5**	**82.8**	**83.6**

Table 2: Performance comparison on the MCEval benchmark. Marks for FF indicate finetuning-free metrics. † indicates our implementation of FLEUR using the same LVLM (LLaVA-Next-72B) as DISCODE for a fair comparison.

Dataset Statistics. Table 1 compares MCEval with other widely-used image captioning evaluation datasets in terms of domain coverage, number of candidate captions, and average caption length (words per caption). As shown, MCEval contains a moderate number of samples for benchmarking performance and covers a broader range of domains compared to existing datasets.

Examples. Figure 2 shows example images, each paired with two candidate captions. As shown, MCEval presents challenging scenarios because it involves images that are more abstract than realistic, accompanied by captions focusing on shapes and visual patterns.

5 Experiments

5.1 Experimental setting

Datasets and performance measures. We conducted extensive experiments to demonstrate the effectiveness of DISCODE on MCEval and four commonly used benchmarks: Flickr8k-Expert (Hodosh, Young, and Hockenmaier 2013), Flickr8k-CF (Hodosh, Young, and Hockenmaier 2013), Composite (Aditya et al. 2015) and Pascal-50S (Vedantam, Zitnick, and Parikh 2015). For MCEval, we report accuracy for each visual domain as well as the mean accuracy. For Flickr8k-Expert, Flickr8k-CF, and Composite, we utilize Kendall's tau-b (τ_b) and tau-c (τ_c) as performance measures. For Pascal-50S, we report mean accuracy over the HC, HI, HM, and MM annotation types.

Baselines. We compare DISCODE with eight competitive baselines on MCEval, including three reference-free metrics: CLIP-S (Hessel et al. 2021), PAC-S (Sarto et al. 2023) and FLEUR (Lee, Park, and Kang 2024), and five reference-based metrics: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Zitnick, and Parikh 2015), and Polos (Wada et al. 2024). FLEUR is the most relevant to our work, as it uses score smoothing.

Implementation details. DISCODE is implemented as described in Section 3.4. For learning-based metrics, we utilize their officially released pre-trained checkpoints.

5.2 Experimental results

Performance on Diverse Domains. Table 2 summarizes the performance comparison on the MCEval benchmark. Overall, DISCODE consistently achieves superior performance across all visual domains, highlighting its robustness to diverse image styles such as paintings and abstract drawings. We observe performance drops in learning-based methods like PAC-S and Polos on non-real domains, likely due to their finetuning targeting real images. By contrast, DISCODE exhibits greater stability across domains, outperforming FLEUR due to its more robust score estimation strategy via the ATT loss minimization.

Performance on Real Image Domain. Table 3 compares DISCODE with state-of-the-art methods on the four real-image benchmark datasets. DISCODE achieves comparable or even better performance than the other methods. On Flickr8k-CF, it outperforms G-VEval, which utilizes a proprietary LVLM (GPT-4o), demonstrating the strong adaptability of DISCODE in real-image evaluation scenarios.

5.3 Analysis

Ablation Study. Table 4 presents the ablation study results examining three key DISCODE components: the cross-entropy term H in Eq. (2), the divergence term D_α in Eq. (2), and the weighting parameter α in Eq. (4). The performance drop resulting from the removal of each component confirms their collective contribution to the model.

Which rating scale is best?. Table 5 compares four different rating scales: (1) a continuous scale from 0.0 to 1.0, (2) a five-point discrete scale from 1 to 5, (3) a ten-point discrete scale from 0 to 9, and (4) a letter-based scale from A to E. We observe that numerical scales perform better, with the continuous scale slightly outperforming discrete scales. This confirms that instructing LVLMs to assign more granular scores is effective in improving correlation with human evaluations. With the continuous scale, the decimal place token always appears immediately before the target digit. This likely helped stabilize the output token probability distribution during autoregressive decoding, resulting in improved performance compared to the ten-point discrete scale.

	Metric	LVLM	Flickr8k-Expert		Flickr8k-CF		Composite		Pascal-50S
			Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c	Accuracy
Reference-based	BLEU-4 (Papineni et al. 2002)		30.6	30.8	16.9	8.7	28.3	30.6	74.0
	ROUGE (Lin 2004)		32.1	32.3	19.9	10.3	30.0	32.4	78.0
	METEOR (Banerjee and Lavie 2005)		41.5	41.8	22.2	11.5	36.0	38.9	81.1
	CIDEr (Vedantam, Zitnick, and Parikh 2015)		43.6	43.9	24.6	12.7	34.9	37.7	80.1
	SPICE (Anderson et al. 2016)		51.7	44.9	24.4	12.0	38.8	40.3	78.7
	RefCLIP-S (Hessel et al. 2021)		52.6	53.0	36.4	18.8	51.2	55.4	83.3
	RefPAC-S++ (ViT-L) (Sarto et al. 2024b)		–	57.9	38.8	–	–	61.6	84.7
	Polos (Wada et al. 2024)		–	56.4	37.8	–	–	57.6	86.5
	DENEB (ViT-L) (Matsuda et al. 2024)		–	56.8	38.3	–	–	58.2	**87.8**
Reference-free	UMIC (Lee et al. 2021)		–	46.8	–	–	–	56.1	85.1
	CLIP-S (Hessel et al. 2021)		51.1	51.2	34.4	17.7	49.8	53.8	80.9
	InfoMetIC+ (Hu et al. 2023)		–	55.5	36.6	–	–	59.3	86.5
	HiFi-Score (Yao, Wang, and Chen 2024)		–	58.4	–	–	–	65.8	83.0
	BRIDGE (ViT-L) (Sarto et al. 2024a)		55.4	55.8	36.3	19.0	52.9	57.2	82.9
	HICE-S (Zeng et al. 2024)		55.9	56.4	37.2	19.2	53.1	57.9	86.1
	PAC-S++ (ViT-L) (Sarto et al. 2024b)		–	57.4	38.5	–	–	62.0	82.4
	CAMScore (Cui et al. 2025)		54.8	55.6	37.5	19.3	53.4	57.5	85.8
	FLEUR (Lee et al. 2024)	✓	–	53.0	38.6	–	–	63.5	83.2
	G-VEval (GPT-4o) (Tong et al. 2025)	✓	**61.5**	**59.7**	38.7	20.2	58.3	63.0	82.3
	FLEUR-LV†	✓	55.3	55.7	40.1	20.7	60.7	65.7	83.8
	DISCODE-LV (Ours)	✓	55.7	56.1	**40.2**	**20.8**	**61.1**	**66.0**	84.5
	FLEUR-IN†	✓	56.5	56.9	36.4	18.8	59.8	64.2	80.8
	DISCODE-IN (Ours)	✓	57.7	58.1	40.1	**20.8**	60.5	64.9	83.5

Table 3: Comparison with state-of-the-art metrics on Flickr8k-Expert, Flickr8k-CF, Composite and Pascal-50S. FLEUR† and DISCODE utilize LLaVA-Next-72B (LV) and InternVL-2.5-78B (IN). Best results are marked in bold, and best results among LVLM-based approaches are underlined.

Method	FEX τ_c	FCF τ_b	Com τ_c	Pascal	MCEval
DISCODE	**56.1**	**40.2**	**66.0**	**84.5**	**83.6**
w/o cross-entropy term H	54.6	39.5	63.1	83.8	81.8
w/o divergence term D_α	49.9	39.9	64.4	83.0	80.9
w/o adaptive definition for α	55.6	**40.2**	65.4	84.3	83.0

Table 4: Ablation study results.

Scale	FEX τ_c	FCF τ_b	Com τ_c	Pascal	MCEval
1 to 5	54.4	40.1	65.1	**84.5**	**83.6**
0 to 9	55.4	40.1	65.8	84.2	83.5
A to E	54.4	39.8	65.7	83.7	83.5
0.0 to 1.0	**56.1**	**40.2**	**66.0**	**84.5**	**83.6**

Table 5: Comparison of rating scales.

Generalization Across LLMs. To investigate the compatibility of DISCODE with different LLMs, we applied it to four different LLMs within the framework of LLaVA-Next: Llama-3-8B (Dubey et al. 2024), Vicuna-13B (Chiang et al. 2023), Nous-Hermes-2-Yi-34B (Research 2024) and Qwen-1.5-72B (Team 2024). Results summarized in Table 6 indicate consistent improvements over raw scores and FLEUR in most cases, verifying the broad applicability of DISCODE. We also see that larger models exhibit higher performance, highlighting the scalability advantages.

Generalization Across LVLMs. To further explore the applicability of DISCODE, we applied it to six leading LVLMs: InternVL-2.5-8B/78B (Chen et al. 2024b,a), Qwen2-VL-7B/72B-Instruct (Bai et al. 2023; Wang et al. 2024a), CogVLM2-Chat-19B (Wang et al. 2024b; Hong et al. 2024), and Mini-CPM-V-2.6 (Hu et al. 2024; Yao et al. 2024). As shown in Table 7, DISCODE consistently improves performance across these models. Among all results in Tables 6 and 7, the best performance is achieved by InternVL-2.5-78B on Flickr8k-Expert (58.1 τ_c), by Qwen-VL-72B-Instruct on Composite (66.7 τ_c) and MCEval (83.8%), and by LLaVA-Next-Qwen-1.5-72B on Flickr8k-CF (40.2 τ_b) and Pascal-50S (84.5%). These results underscore that our approach is effective regardless of the pre-training methods or architectures of the LVLMs.

Generalization Across Divergence Measures. Since the divergence term is a critical component of DISCODE, we examine its performance using different divergence measures. Specifically, we replace the weighted KL divergence with the Jensen–Shannon divergence, the beta divergence, the Rényi divergence, and the standard KL divergence. In Table 8, we observe that DISCODE can effectively leverage these divergence measures. The results also justify our selection of the weighted KL divergence, for which we derived an analytical solution, as the best-performing option.

LLM	Method	FEX τ_c	FCF τ_b	Com τ_c	Pascal	MCEval
LLama 3-8B	Raw score	21.3	23.6	47.5	57.2	52.1
	FLEUR	49.6	33.5	50.4	79.2	76.8
	DISCODE	**51.1**	**36.2**	**61.2**	**81.9**	**77.4**
Vicuna 13B	Raw score	24.5	29.8	57.0	64.4	55.0
	FLEUR	52.1	38.0	61.9	83.5	80.6
	DISCODE	**52.6**	**38.4**	**62.9**	**83.9**	**81.4**
Hermes Yi-34B	Raw score	16.3	22.9	54.6	60.5	61.0
	FLEUR	54.4	**39.5**	**66.1**	83.4	79.7
	DISCODE	**55.0**	**39.5**	**66.1**	**84.1**	**82.5**
Qwen 1.5-72B	Raw score	32.2	35.5	63.1	67.3	61.6
	FLEUR	55.7	40.1	65.7	83.8	82.5
	DISCODE	**56.1**	**40.2**	**66.0**	**84.5**	**83.6**

Table 6: Performance comparison across four LLMs within the LLaVA-NeXT framework.

LVLM	Metric	FEX τ_c	FCF τ_b	Com τ_c	Pascal	MCEval
IntVL 8B	Raw score	26.9	30.4	55.0	66.0	57.9
	FLEUR	56.3	36.7	58.4	76.0	64.5
	DISCODE	**57.4**	**39.2**	**59.6**	**80.2**	**66.5**
IntVL 78B	Raw score	42.0	35.7	59.6	72.8	66.3
	FLEUR	56.9	36.4	64.2	80.8	74.0
	DISCODE	**58.1**	**40.1**	**64.9**	**83.5**	**78.1**
Qwen2 7B-I	Raw score	14.5	39.1	51.7	60.1	53.3
	FLEUR	52.6	**39.5**	52.7	81.5	**75.2**
	DISCODE	**52.9**	**39.6**	**66.2**	**83.3**	**75.2**
Qwen2 72B-I	Raw score	12.8	32.2	63.0	66.9	65.8
	FLEUR	54.1	**40.0**	66.3	83.7	82.4
	DISCODE	**54.4**	**40.0**	**66.7**	**84.1**	**83.8**
CogVL 19B	Raw score	13.7	14.3	30.6	67.3	54.5
	FLEUR	39.1	29.8	44.2	79.6	66.1
	DISCODE	**40.3**	**31.9**	**53.0**	**80.2**	**68.6**
MinC 2.6	Raw score	23.6	31.2	58.7	61.5	58.0
	FLEUR	53.0	**39.7**	**66.2**	83.8	83.0
	DISCODE	**53.5**	**39.7**	**66.2**	**83.8**	**83.3**

Table 7: Performance comparison across six LVLMs.

Divergence	FEX τ_c	FCF τ_b	Com τ_c	Pascal	MCEval
w/o divergence	49.9	39.9	64.4	83.0	80.9
Rényi divergence	55.9	40.1	65.8	84.4	83.0
Beta divergence	55.3	40.0	**66.0**	84.2	81.5
Jensen–Shannon div.	55.6	40.1	65.7	84.1	81.6
Kullback-Leibler div.	55.6	**40.2**	65.4	84.3	83.0
Weighted KLD	**56.1**	**40.2**	**66.0**	**84.5**	**83.6**

Table 8: DISCODE with various divergence measures.

6 Conclusion

We introduced DISCODE, a novel test-time adaptive decoder for LVLM-based image captioning evaluation. By incorporating a unimodal prior distribution into the ATT loss, DISCODE robustly estimates the evaluation score distribution, thereby achieving better alignment with human judgments. We also introduced the MCEval benchmark, consisting of 18,000 image-caption pairs designed to benchmark the robustness of evaluation metrics across multiple visual domains. Our experiments demonstrated the superiority of DISCODE on MCEval and four representative real-image benchmarks, achieving state-of-the-art performance.

Limitations and Future Work. Since our approach leverages latent decoder features, it cannot be applied to proprietary LVLMs like GPT-4o, which do not support feature extraction functionalities. To further enhance performance of open-source LVLMs, extending bias mitigation techniques to tasks beyond image captioning evaluation remains a promising future research direction. We believe our work significantly advances the field of LVLM-based numerical evaluation and provides a solid foundation for future developments from both technical and dataset perspectives.

Acknowledgments

This work was supported by DENSO IT LAB Recognition, Control, and Learning Algorithm Collaborative Research Chair (Science Tokyo). This work was also supported by JSPS KAKENHI Grant Numbers 23H00490 and 25K03135.

References

Aditya, S.; Yang, Y.; Baral, C.; Fermüller, C.; and Aloimonos, Y. 2015. From Images to Sentences through Scene Description Graphs using Commonsense Reasoning and Knowledge. *arXiv preprint arXiv:1511.03292*.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *Proc. European Conference on Computer Vision (ECCV)*, 382–398.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proc. ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Cui, E.; Zhu, J.; Ye, S.; Tian, H.; Liu, Z.; Gu, L.; Wang, X.; Li, Q.; Ren, Y.; Chen, Z.; Luo, J.; Wang, J.; Jiang, T.; Wang, B.; He, C.; Shi, B.; Zhang, X.; Lv, H.; Wang, Y.; Shao, W.; Chu, P.; Tu, Z.; He, T.; Wu, Z.; Deng, H.; Ge, J.; Chen, K.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2024a. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint arXiv:2412.05271*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2024b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24185–24198.

Chiang, W.-L.; Li, Z.; Lin, Z.; Sheng, Y.; Wu, Z.; Zhang, H.; Zheng, L.; Zhuang, S.; Zhuang, Y.; Gonzalez, J. E.; Stoica,

I.; and Xing, E. P. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Cui, T.; Bai, J.; Wang, G.-H.; Chen, Q.-G.; Xu, Z.; Luo, W.; Zhang, K.; and Shi, Y. 2025. Evaluating Image Caption via Cycle-consistent Text-to-Image Generation. *arXiv preprint arXiv:2501.03567*.

Dubey, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Gomes, G. E. C.; Zerva, C.; and Martins, B. 2025. Evaluation of Multilingual Image Captioning: How far can we get with CLIP models? In *Findings Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Hessel, J.; Holtzman, A.; Forbes, M.; Le Bras, R.; and Choi, Y. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7514–7528.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *Journal of Artificial Intelligence Research (JAIR)*, 47: 853–899.

Hong, W.; Wang, W.; Ding, M.; Yu, W.; Lv, Q.; Wang, Y.; Cheng, Y.; Huang, S.; Ji, J.; Xue, Z.; Zhao, L.; Yang, Z.; Gu, X.; Zhang, X.; Feng, G.; Yin, D.; Wang, Z.; Qi, J.; Song, X.; Zhang, P.; Liu, D.; Xu, B.; Li, J.; Dong, Y.; and Tang, J. 2024. CogVLM2: Visual Language Models for Image and Video Understanding. *arXiv preprint arXiv:2408.16500*.

Hu, A.; Chen, S.; Zhang, L.; and Jin, Q. 2023. InfoMetIC: An Informative Metric for Reference-free Image Caption Evaluation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 3171–3185.

Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; Zhang, X.; Thai, Z. L.; Zhang, K.; Wang, C.; Yao, Y.; Zhao, C.; Zhou, J.; Cai, J.; Zhai, Z.; Ding, N.; Jia, C.; Zeng, G.; Li, D.; Liu, Z.; and Sun, M. 2024. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. In *Proc. Conference on Language Modeling (COLM)*.

Jiang, M.; Huang, Q.; Zhang, L.; Wang, X.; Zhang, P.; Gan, Z.; Diesner, J.; and Gao, J. 2019. TIGEr: Text-to-Image Grounding for Image Caption Evaluation. In *Proc. Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Lee, H.; Yoon, S.; Dernoncourt, F.; Bui, T.; and Jung, K. 2021. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 220–226.

Lee, H.; Yoon, S.; Dernoncourt, F.; Kim, D. S.; Bui, T.; and Jung, K. 2020. ViLBERTScore: Evaluating Image Caption Using Vision-and-Language BERT. In *Proc. EMNLP Workshop on Evaluation and Comparison of NLP Systems*, 34–39.

Lee, Y.; Park, I.; and Kang, M. 2024. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 3732–3746.

Li, B.; Zhang, K.; Zhang, H.; Guo, D.; Zhang, R.; Li, F.; Zhang, Y.; Liu, Z.; and Li, C. 2024. LLaVA-NeXT: Stronger LLMs Supercharge Multimodal Capabilities in the Wild. *https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proc. ACL Workshop on Text Summarization Branches Out*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023a. Visual Instruction Tuning. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023b. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2511–2522.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; Sun, Y.; Deng, C.; Xu, H.; Xie, Z.; and Ruan, C. 2024. DeepSeek-VL: Towards Real-World Vision-Language Understanding. arXiv:2403.05525.

Mathew, M.; Bagal, V.; Pérez Tito, R.; Karatzas, D.; Valveny, E.; and Jawahar, C. V. 2022. InfographicVQA. In *Proc. IEEE Conference on Winter Conference on Applications of Computer Vision (WACV)*, 1696–1706.

Matsuda, K.; Wada, Y.; and Sugiura, K. 2024. DENEB: A Hallucination-Robust Automatic Evaluation Metric for Image Captioning. In *Proc. Asian Conference on Computer Vision (ECCV)*, 3570–3586.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Peng, X.; Bai, Q.; Xia, X.; Huang, Z.; Saenko, K.; and Wang, B. 2019. Moment Matching for Multi-Source Domain Adaptation. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 1406–1415.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proc. International Conference on Machine Learning (ICML)*, volume 139, 8748–8763.

Research, N. 2024. Nous Hermes 2. *https://huggingface.co/NousResearch/Nous-Hermes-2-Yi-34B*.

Sarto, S.; Barraco, M.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2023. Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6914–6924.

Sarto, S.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024a. BRIDGE: Bridging Gaps in Image Captioning Evaluation with Stronger Visual Cues. In Leonardis, A.; Ricci, E.; Roth,

S.; Russakovsky, O.; Sattler, T.; and Varol, G., eds., *Proc. European Conference on Computer Vision (ECCV)*, 70–87.

Sarto, S.; Moratelli, N.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024b. Positive-Augmented Contrastive Learning for Vision-and-Language Evaluation and Training. *arXiv preprint arXiv:2410.07336*.

Team, Q. 2024. Introducing Qwen1.5. *https://qwenlm.github.io/ blog/qwen1.5/*.

Tong, T. C.; He, S.; Shao, Z.; and Yeung, D.-Y. 2025. G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o. In *Proc. AAAI Conference on Artificial Intelligence*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

Wada, Y.; Kaneda, K.; Saito, D.; and Sugiura, K. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13559–13568.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Fan, Y.; Dang, K.; Du, M.; Ren, X.; Men, R.; Liu, D.; Zhou, C.; Zhou, J.; and Lin, J. 2024a. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

Wang, S.; Yao, Z.; Wang, R.; Wu, Z.; and Chen, X. 2021. FAIEr: Fidelity and Adequacy Ensured Image Caption Evaluation. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14050–14059.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2024b. CogVLM: Visual Expert for Pretrained Language Models. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E. H.; Le, Q. V.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Yao, Z.; Wang, R.; and Chen, X. 2024. HiFi-Score: Fine-grained Image Description Evaluation with Hierarchical Parsing Graphs. In *Proc. European Conference on Computer Vision (ECCV)*, 441–458.

Yi, Y.; Deng, H.; and Hu, J. 2020. Improving Image Captioning Evaluation by Considering Inter References Variance. In *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*, 985–994.

Yuan, W.; Neubig, G.; and Liu, P. 2021. BARTScore: Evaluating Generated Text as Text Generation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Proc. Annual Conference on Neural Information Processing Systems (NeurIPS)*, 27263–27277.

Zeng, Z.; Sun, J.; Zhang, H.; Wen, T.; Su, Y.; Xie, Y.; Wang, Z.; and Chen, B. 2024. HICEScore: A Hierarchical Metric for Image Captioning Evaluation. In *Proc. ACM International Conference on Multimedia (ACMMM)*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proc. International Conference on Learning Representations (ICLR)*.