G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o

Tony Cheng Tong[1*], **Sirui He**[2*], **Zhiwen Shao**[1,3†], **Dit-Yan Yeung**[1†]
[1]The Hong Kong University of Science and Technology
[2]National University of Singapore
[3]China University of Mining and Technology
ztangaj@connect.ust.hk, he.sirui@u.nus.edu, zhiwen@ust.hk, dyyeung@cse.ust.hk

Abstract

Evaluation metric of visual captioning is important yet not thoroughly explored. Traditional metrics like BLEU, METEOR, CIDEr, and ROUGE often miss semantic depth, while trained metrics such as CLIP-Score, PAC-S, and Polos are limited in zero-shot scenarios. Advanced Language Model-based metrics also struggle with aligning to nuanced human preferences. To address these issues, we introduce G-VEval, a novel metric inspired by G-Eval and powered by the new GPT-4o. G-VEval uses chain-of-thought reasoning in large multimodal models and supports three modes: reference-free, reference-only, and combined, accommodating both video and image inputs. We also propose MSVD-Eval, a new dataset for video captioning evaluation, to establish a more transparent and consistent framework for both human experts and evaluation metrics. It is designed to address the lack of clear criteria in existing datasets by introducing distinct dimensions of Accuracy, Completeness, Conciseness, and Relevance (ACCR). Extensive results show that G-VEval outperforms existing methods in correlation with human annotations, as measured by Kendall tau-b and Kendall tau-c. This provides a flexible solution for diverse captioning tasks and suggests a straightforward yet effective approach for large language models to understand video content, paving the way for advancements in automated captioning.

Code — https://github.com/ztangaj/gveval
Extended version — https://arxiv.org/abs/2412.13647

1 Introduction

Visual captioning, the task of generating descriptive text from visual content, represents a crucial intersection between computer vision and natural language processing. This field primarily addresses the complex challenge of enabling machines to interpret and articulate visual data. Recently, researchers have integrated large language models (LLMs) to enhance the capabilities of visual captioning systems, resulting in more precise and generalizable captions that benefit numerous downstream applications. These advanced systems, known as large vision-language models (LVLMs), now assist the visually impaired, improve educational technologies, and enhance the autonomy of robotics (Li et al. 2020; Zhang et al. 2021).

Candidate: "A dog rolling in the grass ."
Reference Captions: {"A little tan dog with large ears running through the grass .", "A playful dog is running through the grass .", "A small dogs ears stick up as it runs in the grass .", "The small dog is running across the lawn .", "this is a small beige dog running through a grassy field"}

Human	G-VEval (ours)	CLAIR	BLEU-4	METEOR	CIDER
0.6	**0.498**	0.800	0.000	0.203	0.296

Figure 1: Example of caption evaluation.

Despite these advancements, evaluating V-LLMs remains challenging. Traditional metrics like BLEU, METEOR, CIDEr, and ROUGE often miss the semantic depth of captions. Trained metrics such as CLIP-Score and PAC-S, and Polos offer better language understanding but are limited in zero-shot scenarios. Additionally, metrics leveraging LLMs, like CLAIR, show strong human correlation but face limitations in interpretability and applicability to tasks beyond reference-only evaluation.

To address these challenges, we introduce G-VEval, a novel metric inspired by G-Eval (Liu et al. 2023) in natural language generation (NLG). G-Eval pioneered the use of GPT-4 for evaluation by leveraging chain-of-thought reasoning and addressing the probabilistic nature of model outputs through the calculation of expected values. Building on these innovations, G-VEval incorporates genuine chain-of-thought reasoning within large multimodal models and extends this approach to visual content, supporting three evaluation modes: reference-free, reference-only, and combined, making it applicable to both image and short video captioning. Additionally, we propose MSVD-Eval, a new dataset designed to address the limitations of existing evaluation datasets by providing clear criteria for assessing video captions across four dimensions: Accuracy, Completeness, Conciseness, and Relevance (ACCR).

G-VEval aims to bridge the gaps left by existing metrics

*These authors contributed equally.
†Corresponding authors.

like CLAIR and G-Eval by integrating the strengths of both, while also extending the evaluation framework to video captioning. By leveraging in-context chain-of-thought reasoning and multimodal capabilities, G-VEval delivers consistent and high-quality evaluations across a range of captioning tasks. An example of G-VEval evaluation result is shown in Figure 1. Our results demonstrate that G-VEval achieves superior correlation with human judgments compared to existing methods, offering a robust and adaptable evaluation framework for future research in automated captioning.

2 Related Work

In this section, we review existing metrics for evaluating image and video captioning, categorizing them into untrained metrics, trained metrics, and advanced language model-based metrics. Our goal is to highlight the strengths and limitations of each approach, ultimately establishing the need for a more versatile and robust metric like G-VEval.

2.1 Untrained Metrics

Untrained metrics primarily rely on n-gram matching between generated captions and reference captions. These metrics, including BLEU, ROUGE, METEOR, and CIDEr, are popular due to their simplicity and ease of implementation.

BLEU calculates the precision of n-grams in the generated caption against reference captions (Papineni et al. 2002). While widely used in both machine translation and image captioning, BLEU struggles with synonyms and varied sentence structures, which can lead to lower scores for high-quality captions generated by advanced models like V-LLMs. ROUGE focuses on recall by comparing overlapping n-grams, word sequences, and word pairs between the generated and reference captions (Lin 2004). This metric is commonly used in text summarization and image captioning but shares similar limitations with BLEU regarding semantic depth and synonym handling. METEOR combines precision and recall while incorporating synonym matching, stemming, and paraphrase detection (Banerjee and Lavie 2005). It offers a more nuanced evaluation compared to BLEU and ROUGE but still relies heavily on word-level matches.

CIDEr uses TF-IDF weighting for n-grams, emphasizing consensus among multiple references (Vedantam, Zitnick, and Parikh 2015). It is specifically designed for image captioning evaluation but can falter when the generated captions use different wording than the references, especially with synonyms. Despite their widespread use, these untrained metrics often fail to align well with human judgment, particularly for captions generated by models that use advanced language understanding, such as V-LLMs. While these metrics can be applied to video captioning tasks, they do not account for visual content, limiting their effectiveness in this domain.

2.2 Trained Metrics

Trained metrics leverage pre-trained embeddings or human-labeled data, offering greater flexibility in language understanding. Embedding-based metrics, such as BERTScore, evaluate the similarity between generated and reference captions using contextual embeddings from BERT (Zhang et al. 2019). MoverScore enhances BERTScore with soft alignments and advanced aggregation methods (Zhao et al. 2019). These metrics are more flexible with synonyms and sentence segmentation but do not consider visual content, which limits their alignment with human preferences.

To address this gap, researchers have developed metrics based on the cross-modal embeddings of vision-language models. CLIP-Score measures the similarity between generated captions and image content using the CLIP model (Hessel et al. 2021). CLIP-ViT-B-32 and CLIP-ViT-L-14 are commonly used versions, encoding images into 512-dimensional and 768-dimensional vectors, respectively. However, CLIP's encoding lacks the detail needed for fine-grained visual captioning, and its applicability to video captioning is limited. The only notable work in this area is EMScore, which evaluates video captioning via coarse-grained and fine-grained embedding matching (Shi et al. 2022).

Supervised metrics, such as PAC-S and Polos, are trained on datasets derived from human evaluations, showing high correlation with human preferences. PAC-S uses contrastive learning and human-labeled data to evaluate captions, emphasizing positive augmentation (Sarto et al. 2023). Polos, developed using multimodal metric learning from human feedback, is effective in aligning with human judgments (Wada et al. 2024). However, their dependence on training data can lead to weak performance in zero-shot settings, limiting their broader applicability across diverse datasets.

2.3 Advanced Language Model-Based Metrics

Advanced language model-based metrics leverage the capabilities of large language models to provide more robust evaluations. CLAIR is an example of such a metric, using LLMs with simple prompts to evaluate image captions (Chan et al. 2023). While CLAIR shows strong performance in human correlation, it is limited to reference-only evaluation and lacks interpretability due to its reliance on simple prompts. Additionally, CLAIR has not been extended to video captioning tasks, which restricts its broader applicability.

G-Eval, although not a visual captioning evaluation metric, presents a more structured approach to utilizing LLMs for evaluation tasks, specifically in the context of summarization (Liu et al. 2023). Unlike CLAIR, G-Eval calculates the expected value of the output from LLMs such as GPT-3 and GPT-4, addressing the challenges associated with the probabilistic nature of LLMs. While G-Eval claims to use chain-of-thoughts (CoT) reasoning by including evaluation steps in the prompt, it often produces single-digit outputs, lacking a genuine in-context reasoning process, which may limit the effectiveness of CoT.

3 Methodology

G-VEval leverages GPT-4o, a large language model with vision capabilities, to evaluate model performance in image and video captioning tasks. G-VEval provides a framework that generates evaluation scores highly aligned with human

preferences by using prompts. The prompt consists of five modules: 1) Evaluation Criteria; 2) Evaluation Steps: utilizing the Chain-of-Thought (CoT) to guide the LLM in a step-by-step manner, enhancing performance; 3) Score Function: formatting and restricting the output of the LLM; 4) Reference: attaching reference captions from human annotators as ground truth; 5) Original Visual Content: original image or representative frames of the video clip. These modules can be modified to fit the settings of reference-only, reference-free, and combined reference and visual content.

In the following sections, we present selected examples of the prompts used for evaluation. A complete set of all prompts utilized in this study is provided in the appendix for further reference.

3.1 Evaluation Criteria

The evaluation criteria provide clear instructions to define the task of evaluation. The criteria are designed to ensure consistency across different captioning tasks, whether for images or videos.

General Evaluation Criteria for Image and Video Captioning. For both image and video captioning tasks, where an overall score is required, the evaluation criteria are standardized to ensure a uniform approach across different tasks: *Score (from 0 to 100) - selection of important content from the references and the visual content. The generated caption should accurately describe the important aspects of the visual content while including the essential information from the references. Annotators were instructed to penalize captions that contained redundancies and excess information.*

This general approach applies universally to evaluate the overall quality of captions, ensuring that both image and video content are assessed with the same rigor and consistency.

ACCR Evaluation Criteria for Video Captioning. While traditional video captioning metrics provide a single score for overall quality, our framework introduces a more granular approach with the ACCR evaluation criteria. ACCR stands for Accuracy, Completeness, Conciseness, and Relevance, which are the four dimensions used to comprehensively assess the quality of video captions:

- **Accuracy.** Does the caption correctly describe the entities and actions shown in the video without errors or hallucinations?

- **Completeness.** Does the caption cover all significant events and aspects of the video, including dynamic actions and possible scene transitions?

- **Conciseness.** Is the caption clear and succinct, avoiding unnecessary details and repetition?

- **Relevance.** Is the caption pertinent to the video content, without including irrelevant information or questions?

The ACCR dimensions allow for a detailed and multidimensional evaluation of video captions, offering more than just an overall score. By breaking down the evaluation into these four critical areas, ACCR provides a nuanced understanding of caption quality, which is essential for improving the performance of video captioning systems. Additionally, by separating the evaluation dimensions, we reduce the variance in evaluation scores. This approach forces both the evaluation metrics and human annotators to assess captions from the same angles, leading to more consistent and fair evaluations by minimizing inter-human variance.

3.2 Evaluation Steps

The Evaluation Steps leverage the Chain-of-Thought (CoT) approach, guiding GPT-4o to perform the task in a structured, step-by-step manner. This method significantly enhances the performance of the LLM by providing detailed intermediate steps for the evaluation task. According to the Chain-of-Thought paper by Wei et al. (Wei et al. 2022), this technique improves the model's reasoning capabilities.

The example below demonstrates the evaluation steps used for image captioning, generated by GPT-4-Turbo. These steps ensure that the LLM considers all relevant aspects of the content when generating its evaluations. A full set of evaluation steps, including those for video captioning, is provided in the appendix.

Evaluation steps for images.

1: Carefully observe the provided image to understand its main content.
2: Read the reference captions carefully to identify the important information they highlight.
3: Compare the generated caption to both the reference captions and the visual content of the image.
4: Assess how well the generated caption covers the main points of the visual content and the reference captions, and how much irrelevant or redundant information it contains.
5: Assign an integer score from 0 to 100, considering both the alignment with the image and the inclusion of key points from the references.

This structured approach ensures that all relevant aspects of the image and its captions are thoroughly evaluated, contributing to a more accurate and reliable scoring process. For video captioning tasks, similar steps are used, with adaptations to account for the temporal dynamics of video content.

3.3 Score Function

The score function formats and restricts the output of the LLM, ensuring consistency and clarity in the evaluation process. G-VEval supports two scoring settings depending on the specific requirements of the task:

- **Scoring Setting.** In this setting, the score ranges from 0 to 100, providing a fine-grained evaluation scale. This approach is particularly useful when a more detailed assessment is needed, allowing for a broader range of possible scores.

- **Rating Setting.** Alternatively, the score can be restricted to an integer between 1 and 5, offering a simpler and more straightforward evaluation. This setting can be beneficial in scenarios where a coarser granularity is sufficient.

For the purposes of this paper, we primarily utilize the scoring setting (0 to 100), as it has demonstrated superior performance in terms of human correlation in our preliminary experiments.

Response Format: You should first give a detailed reason for your score, ending with a sentence like this: The final score is $\{\{score\}\}$. Note that the score should be an integer from 0 to 100, and should be wrapped in dollar signs ($).

Unlike G-Eval, where only a final score is provided, we ensure that GPT-4o outputs a detailed reason for the score, incorporating in-context reasoning. This method enhances the interpretability of the results, as confirmed by our ablation study, which is discussed in the experiment section.

3.4 Handling Probabilistic Outputs

G-VEval handles probabilistic outputs by calculating the expected value of the score using the log probabilities (logprobs) provided by GPT-4o. GPT-4o generates text in an autoregressive manner, where each token's probability distribution is conditioned on the previously generated tokens. This process allows us to derive the expected score by considering the probabilistic distribution over possible outputs.

The expected score, denoted as $E(s)$, is calculated using the formula:

$$E(s) = \sum_{i=1}^{m} i \times p(i), \tag{1}$$

where m represents the maximum possible score, and $p(i)$ is the probability of each score i. The probability $p(i)$ is computed as

$$p(i) = \sum_{j=1}^{n} p(i|R_j) \times p(R_j). \tag{2}$$

Here $p(i|R_j)$ is the probability of score i given the reason R_j, and $p(R_j)$ is the probability of the reason R_j, with n being the total number of possible reasons.

The expected score can then be expressed as

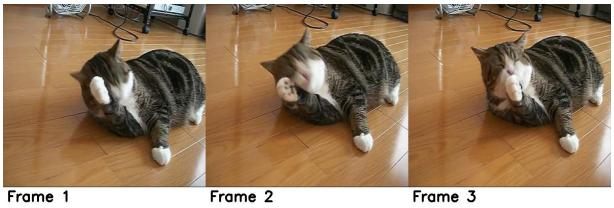
$$E(s) = E_{R_j}(E_s(s|R_j)). \tag{3}$$

This formula highlights that $E(s)$ can be approximated by $E_s(s|R_j)$ when the variance of $E_s(s|R_j)$ is close to zero, as demonstrated in our experimental results. The detailed derivation of these equations is provided in the appendix.

3.5 Reference Captions and Original Visual Content

In G-VEval, reference captions are critical for providing ground truth against which generated captions are evaluated. These captions, provided by different annotators, often emphasize different aspects of the visual content. Therefore, for both image and video captioning tasks, we integrate all reference captions associated with the same visual input to form a comprehensive ground truth. This integrated version ensures that no significant detail is overlooked during evaluation.

For the reference-free and combined-reference settings, visual content is included in the prompt. For image captioning tasks, the original image is directly uploaded to GPT-4o for evaluation. However, video captioning presents unique challenges since GPT-4o does not support direct video encoding. To address this, we sample three frames from each video clip: the first frame, the last frame, and the frame located at the midpoint of the video. These frames are then combined into a single 1536x512 image, with each frame labeled to indicate its position in the sequence (Frame 1, Frame 2, Frame 3), as shown in Figure 2. This setup aids the model's spatial and temporal understanding, allowing it to interpret the sequence of events across frames effectively.



Reference Captions: {"a cat is laying down licking its paw", "a cat is cleaning itself", "a grey cat lying on a wooden floor is rubbing his face with his right fore paw and thereafter licks it", "a large cat is cleaning itself by licking a paw and rubbing the paw over his face"}

Figure 2: Sample frames with annotation of order combined into a single image for video caption evaluation.

The image size ensures GPT-4o's vision encoder processes each frame as a 512x512 tile, leveraging its OCR capabilities for effective positional context.

When adapting G-VEval to video captioning tasks, the evaluation steps are modified to account for the combined frames. This method ensures that G-VEval provides accurate and stable evaluations, making it an effective and versatile tool for both image and video captioning tasks.

4 Experiments

In this section, we evaluate the performance of our G-VEval metric compared to other metrics in the tasks of image and video captioning. We first conduct pre-experiments to establish the optimal settings for G-VEval and then test on the Flickr8k-Expert and Flickr8k-CF datasets for image captioning, as well as the VATEX-EVAL and MSVD-Eval datasets for video captioning. Additionally, we conduct an ablation study to assess the impact of different prompt components on the performance of G-VEval.

4.1 Pre-Experiments

The pre-experiments aim to compare the effectiveness of two settings in G-VEval: the scoring setting (0 to 100) and the rating setting (1 to 5). This comparison helps determine which setting provides better alignment with human judgment and more reliable results.

We conducted experiments on the Flickr8k-Expert dataset, applying both the scoring and rating settings. For each setting, we calculated the variance of $E_s(s|R_j)$ and measured the correlation with human judgments using Kendall's tau-b and tau-c metrics. The observed variance of $E_s(s|R_j)$ for both settings is low (0.014 for scoring and 0.0087 for rating), indicating consistent outputs.

These low variances in both settings suggest that $E(s) = E_{R_j}(E_s(s|R_j))$ can be effectively approximated by $E_s(s|R_j)$. This approximation forms the foundation for our experimental approach, as it allows us to calculate $E_s(s|R_j)$

Setting	Variance	Kendall τ_b	Kendall τ_c
G-VEval-rating	0.0087	54.393	52.468
G-VEval-scoring	0.0144	**60.385**	**58.598**

Table 1: Comparison of G-VEval settings.

directly in subsequent experiments. A detailed proof of this approximation will be provided in the appendix.

Despite the small variance in both settings, the scoring setting significantly outperformed the rating setting in terms of correlation with human judgments, as shown in Table 1. This indicates that the scoring setting offers much better alignment with human judgment. Therefore, we adopt the scoring setting for all subsequent experiments. The G-VEval score is given as $E_s(s|R_j)$.

These findings validate the use of the scoring setting in G-VEval, with its finer granularity allowing for more precise alignment with human preferences, thereby demonstrating superior performance.

4.2 Image Captioning Performance

We tested our G-VEval metric against other metrics on the Flickr8k-Expert and Flickr8k-CF datasets (Hodosh, Young, and Hockenmaier 2013). These datasets are described in detail below.

- **Flickr8k-Expert** consists of 8,000 images, each annotated by three experts with five captions. This dataset provides high-quality reference captions for evaluating captioning models.

- **Flickr8k-CF** (CrowdFlower) includes the same 8,000 images as Flickr8k-Expert but with captions annotated by crowdworkers. This dataset offers a different perspective on captioning quality due to the varied skill levels of annotators.

Table 2 shows the performance of various metrics on the Flickr8k-Expert and Flickr8k-CF datasets. Among the three G-VEval settings, the ref-free setting consistently achieves the highest human judgment correlation scores, establishing a new state-of-the-art. This can be attributed to its ability to evaluate captions purely based on visual content, avoiding potential biases from incomplete or misleading reference captions.

The combined setting, which integrates both reference captions and the original image, performs relatively worse than the ref-free setting. As illustrated in Figure 3, this occurs when reference captions fail to capture critical aspects like the "forest" setting, introducing noise and slightly lowering the correlation score. Nevertheless, the combined setting still achieves high scores, though it underscores the importance of visual content in evaluation.

G-VEval's ref-only setting, relying solely on reference captions, performs lower than the combined setting, emphasizing the need for visual context in accurate evaluations.

Moreover, G-VEval outperforms CLAIR, particularly in Kendall τ_c, indicating that G-VEval is less prone to extreme scores, resulting in more stable and consistent evaluations. This is further demonstrated in Figure 1, where G-VEval delivers a more balanced and accurate evaluation compared to CLAIR, which tends to give extreme scores.

Candidate: "A man scales a rock in the forest."

Reference Captions: {"a man in a harness climbing a rock wall", "A man with a white shirt and helmet is climbing a mountain .", "A man in a white shirt and helmet is using climbing equipment .", "a young man wearing a white helmet climbing up a rock wall", "Man rock climbing looking up the rock"}

Human	G-VEval-Ref-only	G-VEval-Ref-free	G-VEval-Combined
0.8	0.563	**0.830**	0.66

Figure 3: Example of caption evaluation.

4.3 Video Captioning Performance

We evaluate G-VEval on the VATEX-EVAL and MSVD-Eval datasets to assess its effectiveness in video captioning tasks. The evaluation settings for each dataset are described below.

- **VATEX-EVAL** is a comprehensive dataset proposed by EMScore (Shi et al. 2022). It includes diverse video clips with corresponding captions, allowing for robust evaluation of captioning quality. In this evaluation, we use three settings: No Ref, 1 Ref, and 9 Refs. The No Ref setting corresponds to G-VEval-ref-free, while the 1 Ref and 9 Refs settings correspond to our combined evaluation setting, where both the reference captions and visual content are used to generate the scores.

- **MSVD-Eval** is our newly proposed dataset, created to enhance the evaluation of video captioning systems. It consists of 150 video clips selected from the MSVD validation set (Chen and Dolan 2011), with candidate captions generated by Video-LLaMA (Zhang, Li, and Bing 2023). These captions were selected to include both typical failure cases and acceptable captions of LLM-generated video captions. Experts evaluated these captions across four key ACCR dimensions: Accuracy, Completeness, Conciseness, and Relevance, ensuring a comprehensive and detailed assessment framework. For comparison with other metrics, we also provide an overall score (Avg.) by averaging the ACCR scores.

Table 3 shows the performance of various metrics on the VATEX-EVAL dataset under the No Ref, 1 Ref, and 9 Refs settings. Table 4 presents the overall human judgment correlation scores for various metrics on the MSVD-Eval dataset, compared with the averaged human scores. Meanwhile, Table 5 details the performance of G-VEval across the individual ACCR dimensions.

The results from VATEX-EVAL indicate that G-VEval's combined setting, which leverages both reference captions and visual content, provides a substantial improvement over traditional metrics, particularly in the 9 Refs setting. In the MSVD-Eval dataset, G-VEval achieves the highest human judgment correlation across all ACCR dimensions, further validating its capability to handle the nuanced evaluation of both typical failures and high-quality outputs in LLM-generated video captions.

The ACCR dimensions offer a significant advantage by focusing human experts on specific aspects of caption qual-

Metric	Reference Caption	Image Used	Flickr8k-Expert		Flickr8k-CF	
			Kendall τ_b	Kendall τ_c	Kendall τ_b	Kendall τ_c
BLEU-1 (Papineni et al. 2002)	✓		32.2	32.3	17.9	9.3
BLEU-4 (Papineni et al. 2002)	✓		30.6	30.8	16.9	8.7
ROUGE (Lin 2004)	✓		32.1	32.3	19.9	10.3
METEOR (Banerjee and Lavie 2005)	✓		41.5	41.8	22.2	11.5
CIDEr (Vedantam, Zitnick, and Parikh 2015)	✓		43.6	43.9	24.6	12.7
SPICE (Anderson et al. 2016)	✓		51.7	44.9	24.4	12.0
BERT-S (Zhang et al. 2019)	✓		-	39.2	22.8	-
LEIC (Cui et al. 2018)	✓	✓	46.6	-	29.5	-
BERT-S++ (Zhang et al. 2020)	✓		-	46.7	-	-
UMIC (Lee et al. 2021)	✓		-	46.8	-	-
TIGEr (Jiang et al. 2019)	✓	✓	-	49.3	-	-
ViLBERTScore (Lee et al. 2020)	✓	✓	-	50.1	-	-
MID (Huang et al. 2019)	✓		-	54.9	37.3	-
CLIP-S (Hessel et al. 2021)		✓	51.1	51.2	34.4	17.7
PAC-S (Sarto et al. 2023)		✓	53.9	54.3	36.0	18.6
RefCLIP-S (Hessel et al. 2021)	✓	✓	52.6	53.0	36.4	18.8
RefPAC-S (Sarto et al. 2023)	✓	✓	55.5	55.9	37.6	19.5
Polos (Wada et al. 2024)	✓	✓	56.4	-	37.8	-
CLAIR (Chan et al. 2023)	✓		58.3	48.8	38.2	17.0
G-VEval-ref-only	✓		60.4	58.6	37.2	19.4
G-VEval-ref-free		✓	**61.5**	**59.7**	**38.7**	**20.2**
G-VEval-combined	✓	✓	60.5	58.7	38.2	19.9

Table 2: Human judgment correlation scores on Flickr8k-Expert and Flickr8k-CF. The columns "Reference Caption" and "Image Used" indicate whether the metric uses reference captions and/or the original image for evaluation.

Metric	VATEX-EVAL		
	No Ref	1 Ref	9 Refs
BLEU-1	-	12.2	28.9
BLEU-4	-	12.6	22.4
ROUGE	-	12.5	23.8
METEOR	-	16.4	27.6
CIDEr	-	17.3	27.8
BERT-S	-	18.2	29.3
BERT-S++	-	15.2	24.4
EMScore	23.2	28.6	36.8
PAC-S/RefPAC-S	25.1	32.6	31.4
CLAIR	-	36.0	34.8
G-VEval	**39.4**	**44.9**	**48.1**

Table 3: Human judgment correlation scores on VATEX-EVAL dataset.

Metric	MSVD-Eval	
	Kendall τ_b	Kendall τ_c
BLEU-1	40.7	41.2
BLEU-4	34.0	34.4
ROUGE	39.8	40.2
METEOR	45.4	45.9
CIDER	37.3	37.7
EMScore	35.3	35.7
EMScore_ref	50.7	51.3
PAC-S	34.5	34.9
RefPAC-S	52.2	52.8
CLAIR	44.6	40.3
G-VEval-ref-only	60.4	61.1
G-VEval-ref-free	59.6	60.3
G-VEval-combined	**62.9**	**63.7**

Table 4: Human judgment correlation scores on MSVD-Eval dataset (Average Scores).

ity, thereby reducing inter-human variance and improving the reliability of the evaluation. This structured approach allows G-VEval to better align with human judgments, particularly in video captioning tasks where capturing nuances in content is critical.

It is important to note that G-VEval's current design focuses on short-form videos (under 10 seconds), such as those in the MSVD dataset. For longer videos, additional techniques, such as scene detection to divide videos into shorter clips, may be necessary for effective evaluation.

4.4 Ablation Study

To understand the impact of different components of our G-VEval prompt, we conducted an ablation study on the Flickr8k-Expert dataset under the reference-only setting. This study examines how the performance of G-VEval changes when certain key elements of the evaluation process are removed or altered. Specifically, we tested the following settings:

- **G-VEval (full setting).** The original prompt with Chain-of-Thought (CoT) evaluation steps, in-context reasoning, and expected score calculation.

- **G-VEval w/o expected score.** In this setting, instead of calculating the expected value $E(s|R_j)$ from the probabilistic outputs of GPT-4o, we directly use the single score s provided by GPT-4o without considering the probabilistic distribution of possible scores.

- **G-VEval w/o CoT prompt.** This setting removes the

Metric	Acc.	Com.	Con.	Rel.
G-VEval-ref-only	60.4	54.2	55.2	52.5
G-VEval-ref-free	55.3	49.7	**62.2**	**53.4**
G-VEval-combined	**61.4**	**57.6**	58.5	53.0

Table 5: Human judgment correlation scores in Kendall τ_b on MSVD-Eval dataset across ACCR dimensions.

Setting	Kendall τ_b	Kendall τ_c
Full setting	**60.385**	**58.598**
W/o expected score	59.118	54.847
W/o CoT prompt	50.157	48.408
W/o reason in response	52.436	26.944

Table 6: Ablation study results on Flickr8k-Expert.

CoT evaluation steps from the prompt, testing the impact of losing the guided, step-by-step reasoning process.

- **G-VEval w/o reason in response.** Here, we omit the requirement for GPT-4o to provide a detailed reason for the score. The model simply outputs a score, and this score is used without the additional reasoning context.

Table 6 shows that the full G-VEval setting, which includes all components, provides the highest correlation with human judgments. Removing the expected score calculation and using the direct score s slightly reduces performance, indicating the importance of probabilistic handling in score determination. The absence of CoT steps results in a notable drop in Kendall's tau-b and tau-c scores, emphasizing the value of structured, step-by-step reasoning. Lastly, omitting the reason in the response causes a significant decline in Kendall's tau-c, highlighting how critical in-context reasoning is for capturing the nuances of human judgment.

Overall, these results confirm that each component of the G-VEval framework contributes to its effectiveness. The integration of reference captions, visual content, and a structured evaluation approach allows G-VEval to outperform existing metrics in both image and video captioning tasks.

5 Discussion

G-VEval leverages the advanced capabilities of GPT-4o, particularly in language understanding and visual content interpretation, through the use of Chain-of-Thought (CoT) prompting. This approach allows G-VEval to effectively utilize a large multimodal pre-trained dataset and a transformer model with billions of parameters. G-VEval comprehends visual content from multiple perspectives provided by reference captions, enabling a deeper evaluation of candidate captions by comparing their meaning with the visual content. This deeper level of understanding allows G-VEval to outperform traditional n-gram matching methods in aligning with human preferences.

When compared to metrics that use pre-trained embeddings, such as BERTScore (Zhang et al. 2019), G-VEval benefits from GPT-4o's comprehensive embeddings for both language and visual content. Although the exact details of GPT-4o's multimodal embedding model are not publicly available, it is likely influenced by models like BLIP-2's Q-former (Li et al. 2023), which achieves performance comparable to the GPT-4 series. Unlike CLIP-based embeddings used in some metrics (Hessel et al. 2021), GPT-4o's embeddings, potentially enhanced by EVA_CLIP, capture more detailed representations of visual content. The CoT prompting technique further leverages these detailed visual representations, allowing the model to focus on specific image regions, thereby mimicking human-like processing in visual captioning tasks.

G-VEval also performs competitively with training-based metrics such as PAC-S (Sarto et al. 2023). Despite not being fine-tuned specifically for visual captioning evaluation, G-VEval's extensive pre-training, powerful model architecture, and CoT prompting enable it to perform effectively in zero-shot settings, aligning closely with human preferences. Furthermore, G-VEval's adaptability across various tasks is noteworthy; by modifying the prompt, it can be tailored to different evaluation contexts, highlighting its potential for broad applicability in future research.

However, G-VEval has certain limitations. One major drawback is its cost. While GPT-4o is relatively more affordable than some alternatives, it remains more expensive than other metrics due to the token-based pricing model. Another potential concern is the consistency of scoring over time. Although current results demonstrate consistent performance, future updates to the GPT-4o model or changes in prompts could affect this consistency.

Additionally, G-VEval is currently designed for short-form videos, where representative frames effectively capture temporal context. For longer videos, extensions such as scene detection may be necessary to maintain performance.

6 Conclusion

We introduced G-VEval, an innovative evaluation metric designed for image and video caption evaluation. G-VEval harnesses the deep multimodal understanding capabilities of GPT-4o, utilizing the Chain-of-Thought reasoning and expected score calculations based on decoding probability distributions. This metric supports three evaluation modes and excels in scenarios where n-gram and embedding-based metrics fall short, particularly in zero-shot and reference-free contexts. Through extensive experiments, G-VEval has demonstrated state-of-the-art performance and achieved the superior correlation with human evaluations compared to established metrics.

The introduction of MSVD-Eval further enriches the landscape of video caption evaluation by offering a dataset that emphasizes multi-dimensional assessment criteria through the ACCR framework, focusing on Accuracy, Completeness, Conciseness, and Relevance. This approach significantly enhances the reliability and consistency of the evaluation process by focusing human expert assessments on the same aspects.

Looking ahead, we aim to develop an online benchmark platform based on G-VEval, where researchers can evaluate their image and video captioning models, furthering research and innovation in automated visual understanding.

Acknowledgments

This work has been made possible by a Research Impact Fund project (R6003-21) and an Innovation and Technology Fund project (ITS/004/21FP) funded by the Hong Kong Government.

References

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, 382–398. Springer.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72.

Chan, D. M.; Petryk, S.; Gonzalez, J. E.; Darrell, T.; and Canny, J. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Singapore, Singapore: Association for Computational Linguistics.

Chen, D.; and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, 190–200.

Cui, Y.; Yang, G.; Veit, A.; Huang, X.; and Belongie, S. 2018. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5804–5812.

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., *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 7514–7528. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

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*, 47: 853–899.

Huang, X.; Veit, A.; Huang, Z. A.; and Belongie, S. 2019. Learning to evaluate image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12): 2959–2972.

Jiang, Z.; Gan, C.; Wu, J.; Zhao, H.; and Xie, L. 2019. TIGEr: Text-to-Image Grounded Evaluator for Image Captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4281–4290.

Lee, C. Y.; Yoon, J.; Dernoncourt, F.; Bui, T.; and Jung, K. 2021. UMIC: Unreferenced Metric for Image Captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 963–977.

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 *Proceedings of the Workshop on Evaluation and Comparison of NLP Systems*, 34–39.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 19730–19742. PMLR.

Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; Choi, Y.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *Proceedings of the European Conference on Computer Vision*, 121–137.

Lin, C.-Y. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 74–81.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 311–318.

Sarto, S.; Barraco, M.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6914–6924.

Shi, Y.; Yang, X.; Xu, H.; Yuan, C.; Li, B.; Hu, W.; and Zha, Z. 2022. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*.

Vedantam, R.; Zitnick, C. L.; and Parikh, D. 2015. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4566–4575.

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

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the Advances in Neural Information Processing Systems*, 35: 24824–24837.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. BERTScore: Evaluating Text Generation with BERT. *ArXiv*, abs/1904.09675.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. Revisiting BERTScore for Image Captioning: Extensions and Relevance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 834–842.

Zhao, W.; Peyrard, M.; Liu, F.; Gao, Y.; Meyer, C. M.; and Eger, S. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, 563–578.