VC-Inspector: Advancing Reference-free Evaluation of Video Captions with Factual Analysis

Shubhashis Roy Dipta [1*] **Tz-Ying Wu** [2*] **Subarna Tripathi** [2]
[1]**University of Maryland, Baltimore County** [2]**Intel**
sroydip1@umbc.edu tz-ying.wu@intel.com

Abstract

We propose VC-Inspector, a lightweight, open-source large multimodal model (LMM) for *reference-free* evaluation of video captions, with a focus on factual accuracy. Unlike existing metrics that suffer from limited context handling, weak factuality assessment, or reliance on proprietary services, VC-Inspector offers a reproducible and fact-aware alternative that aligns closely with human judgments. To enable robust training and interpretable evaluation, we introduce a systematic framework for generating captions with controllable factual errors, paired with graded quality scores and explanatory annotations. Experiments demonstrate that VC-Inspector achieves state-of-the-art correlation with human judgments, generalizing across diverse domains (e.g., VATEX-Eval, Flickr8K-Expert, and Flickr8K-CF benchmarks) and revealing the potential for caption improvement. Project page: https://dipta007.github.io/VC-Inspector.

Figure 1: Existing *reference-free* metrics like EMScore (Shi et al., 2021) often fail to detect factual inaccuracies and lack a consistent scoring scale. VC-Inspector addresses these limitations by providing *factually grounded*, **interpretable evaluations**.

1 Introduction

Video captions summarize salient objects and actions in videos, supporting downstream applications such as question answering, event localization, and retrieval (P.J. and Kovoor, 2024; Qian et al., 2024; Gabeur et al., 2020). Most evaluation metrics rely on *reference-based* protocols (Papineni et al., 2002; Banerjee and Lavie, 2005; Lin, 2004; Vedantam et al., 2015; Zhang et al., 2020), which compare captions to human-written references. These approaches are costly to scale and often fail to capture semantic equivalence.

Evaluating captions for in-the-wild videos requires *reference-free* protocols that do not rely on ground truth captions, yet this remains underexplored. Existing metrics (Shi et al., 2021; Sarto et al., 2023) typically measure visual-language alignment using pretrained multimodal embeddings (Radford et al., 2021), but are limited by text encoder context length, and lack a consistent scoring scale, making interpretation difficult (Fig. 1). Recently, large proprietary models such as GPT-4o are used to rate captions on a fixed scale (Tong et al., 2024), but these approaches depend heavily on prompt engineering and are not fully reproducible. Additionally, most prior methods are image-centric, rendering them suboptimal for video content.

In this work, we aim to develop a *reference-free* evaluation metric for video captions that removes dependency on human-annotated references while remaining robust and intuitive for humans. Our approach is grounded in **factual accuracy**, as factual elements like objects and actions are critical for video understanding. Ideally, a reliable metric should degrade scores proportionally to factual errors. For example, for a video showing a little girl

[*]These authors contributed equally to this work.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 33657–33672

sitting on a chair, the caption "A little girl is sleeping on a chair" should receive a higher score than "A woman is sleeping on a chair", because the latter deviates more from the video content. However, existing metrics like EMScore (Shi et al., 2021) often fail to capture even basic factual inaccuracies, such as incorrect objects or actions.

To address these limitations, we propose `VC-Inspector`, built on top of a lightweight, open-source large multimodal model (LMM) trained to assess caption quality based on factual correctness. Unlike previous metrics that provide only a score, `VC-Inspector` also generates explanations for its judgments, improving interpretability. One key challenge is the lack of captions with varying degrees of factual quality for training. We overcome this by introducing a novel data generation framework powered by a large language model (LLM), which **controllably** modifies factual elements in ground-truth captions from ActivityNet-Captions (Krishna et al., 2017). This process yields `ActivityNet-FG-It`, a dataset of 44K instances for instruction tuning. Fig. 2 illustrates our data generation and training framework.

We comprehensively evaluate the quality of `VC-Inspector` across multiple complementary settings. We first examine its consistency and reliability on two synthetic datasets, `ActivityNet-FG-Eval` and `YouCook2-FG-Eval`, which contain captions of diverse quality generated using the same pipeline as our instruction-tuning data. Results show that `VC-Inspector` produces stable quality estimates across visual domains and caption lengths. We further assess its correlation with human judgments on two video caption datasets, VATEX-Eval (Shi et al., 2021) and YouCook2-Eval, and extend the evaluation to image caption benchmarks, Flickr8K-Expert and Flickr8K-CF (Hodosh et al., 2013), by viewing images as an extreme case of short videos. Across all datasets, `VC-Inspector` outperforms prior *reference-free* metrics and even surpasses most *reference-based* metrics, highlighting its strong generalization capability. In addition, we evaluate `VC-Inspector` on two hallucination benchmarks, FOIL-COCO (Shekhar et al., 2017) and ActivityNet-FOIL (Shi et al., 2021), where FOIL captions contain object errors. Explicitly trained for object and action grounding, `VC-Inspector` performs strongly on both benchmarks. Beyond score prediction, we further show that the explanatory outputs of `VC-Inspector` not only enhance quality estimation but also enable caption refinement.

To summarize, this work makes the following contributions:

- We introduce a scalable pipeline for synthesizing video captions with controllable factual errors, enabling large-scale training and evaluation, without costly human annotation.
- We propose `VC-Inspector`, a *fact-aware*, *reference-free* video caption evaluator that jointly predicts quality scores and factual error explanations, enhancing interpretability and guiding caption improvement.
- We demonstrate that `VC-Inspector` achieves high correlations with human judgments, outperforming existing *reference-free* methods and rivaling *reference-based* metrics across video and image caption benchmarks, showing strong cross-domain generalization.

2 Related Works

Text-only metrics based on references. Traditional rule-based metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) primarily capture syntactic similarity and fail to reflect semantic meaning. SPICE (Anderson et al., 2016; Li et al., 2023) introduces semantic parsing of objects, attributes, and relations for both reference and candidate captions, but struggle with paraphrases. This has motivated the embedding-based methods, such as BERTScore (Zhang et al., 2020) and its variant (Yi et al., 2020), which consider the intrinsic variance between multiple ground truth captions. More recently, CLAIR (Chan et al., 2023) explores an LLM-as-a-Judge approach for image caption evaluation, demonstrating stronger correlation with human judgments than the above metrics. However, all these metrics depend on reference captions for evaluation and ignore the visual content.

Image-augmented metrics. To overcome the limitation of text-only metrics, image-aware approaches align captions with visual input. VIFIDEL (Madhyastha et al., 2019) matches object names between images and captions, but is restricted to discrete categories and cannot capture motion dynamics in videos. Visual-language model

(VLM) based metrics offer richer semantic alignment through continuous representations. ViLBERTScore (Lee et al., 2020) extends BERTScore using ViLBERT (Lu et al., 2019), but still requires comparison to reference captions. UMIC (Lee et al., 2021) and CLIP-based metrics (Hessel et al., 2021; Jiang et al., 2019; Cui et al., 2018; Sarto et al., 2023; Wada et al., 2024; Zeng et al., 2024; Lee et al., 2024; Shi et al., 2021; Liu and Wan, 2023) relax this constraint by directly modeling image-caption semantic alignment via contrastive learning, though remain limited by the context length of their text encoders. Recently, Maeda et al. (2024) compare the candidate caption to the VLM-derived visual context with a LLM. However, these methods focus on static images and do not model the temporal dynamics, which limits their effectiveness for video caption evaluation.

In summary, text-based methods require costly, high-quality reference captions, while image-based extensions remain less effective for video content. These challenges underscore the need for *reference-free* video caption evaluation tailored to video inputs, a topic that is relatively less studied compared to text-based and image-caption evaluation. EMScore (Shi et al., 2021) was an early attempt that supports evaluating video captions without a reference. Although it considers both frame-level and video-level embeddings, they are derived from an image-based encoder (Radford et al., 2021), and its text encoder is constrained by short context length. PAC-S (Sarto et al., 2023) and FactVC (Liu and Wan, 2023) augment EMScore with positive and negative data synthesis, respectively. While sharing some similarity to this work, they consider only a single level of corruption, with binary (positive/negative) differentiation, whereas our method incorporates captions with varying degrees of quality, enabling a more nuanced evaluation. In addition, these metrics produce a single scalar score without offering explanations of their judgments, posing challenges for interpreting the quality assessment. More recently, G-VEval (Tong et al., 2024) extends G-Eval (Liu et al., 2023b) to video by stitching together three frames per video. However, it is unclear whether this image-based prompting generalizes to longer or more dynamic videos. Moreover, the dependence on large proprietary models such as GPT-4o limits its scalability and reproducibility for widespread use. In contrast, `VC-Inspector` is explicitly trained for video-level factual grounding, produces interpretable explanations, and avoids reliance on proprietary large models, addressing key limitations of prior work.

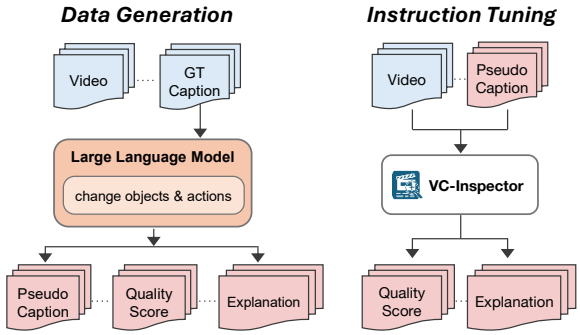


Figure 2: (left) We present a data generation pipeline designed to systematically create synthetic video captions with diverse quality scores, along with explanations for the assigned scores. (right) This dataset was subsequently used for instruction tuning the `VC-Inspector`.

3 Methodology

3.1 Overview

Video caption quality estimation aims to quantify the correctness of a caption $\hat{X} = \mathcal{M}(V)$ given a video V, where $\mathcal{M}$ denotes a video captioning model. Prior work falls into two categories: *reference-based* and *reference-free* metrics. The former compares the candidate caption $\hat{X}$ against the ground truth caption X, whereas the latter eliminates this dependency.

We focus on the *reference-free* setting, evaluating $(V, \hat{X})$ without X, as assuming reference captions are always available is impractical. Existing metrics in this category typically leverage pretrained visual-language embeddings (Radford et al., 2021) to assess the semantic alignment between V and $\hat{X}$. However, these methods are limited by the text encoder's context length and lack an interpretable scoring scale. On the contrary, recent work (Tong et al., 2024) employs large proprietary models (e.g., GPT-4o) to support longer contexts, but these approaches rely heavily on prompt engineering and are not fully reproducible.

We argue that a reliable evaluation protocol should be **factually grounded, interpretable, and scalable**. It must reflect the correctness of factual elements (e.g., objects and actions) in the caption relative to the video. For example, captions with missing objects or incorrect actions should receive lower scores, and those with multiple errors should be penalized more severely. Furthermore, evaluation should be intuitive for humans and repro-

ducible for broad adoption.

To address these limitations, we propose VC-Inspector, which leverages a lightweight, open-source large multimodal model (LMM) as the backbone for long-context reasoning and generalized video feature extraction. Our hypothesis is that the strong visual-language reasoning capability of LMMs can transfer effectively to this task. However, pretrained LMMs do not inherently produce factually grounded assessments and therefore require targeted supervision (Table 2). A key challenge is the scarcity of annotated captions exhibiting diverse degrees of factual accuracy for effective training. To overcome this, we introduce a novel data generation pipeline powered by large language models (LLMs) that systematically synthesizes candidate captions paired with graded quality scores and factual error explanations. In this work, we focus explicitly on object and action inaccuracies, which constitute the predominant sources of factual errors in video captions. The overall data generation and training framework is illustrated in Fig. 2.

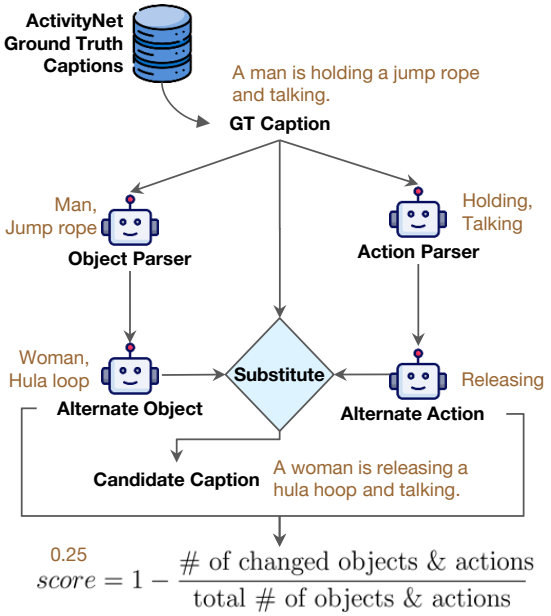


Figure 3: Data generation pipeline to create a synthetic dataset for training VC-Inspector. While both "talking" and "holding" were identified as actions, only "holding" was sampled for replacement in the synthetic dataset.

3.2 Data Generation

To create captions with controlled factual errors, we employ Llama-3.3-70B-Instruct to systematically alter objects and actions in ground truth captions from a supervised video caption dataset. Using this approach, we construct ActivityNet-FG-It, an instruction-tuning dataset for factual grounding derived from the ActivityNet-Captions (Krishna et al., 2017) training set. The data generation pipeline is illustrated in Fig. 3.

Caption generation. Given a ground truth caption X, we first prompt the LLM to extract the set of objects $\mathcal{O} = \{o_1, ..., o_M\}$ and actions $\mathcal{A} = \{a_1, ..., a_N\}$. We then randomly sample $K \sim Unif(0, M)$ objects and $L \sim Unif(0, N)$ actions to replace, forming a subset $\mathcal{R} \subseteq \mathcal{O} \cup \mathcal{A}$, where $|\mathcal{R}| = K + L$ and $Unif(a, b)$ denotes a discrete uniform distribution over integers $\{a, ..., b\}$.

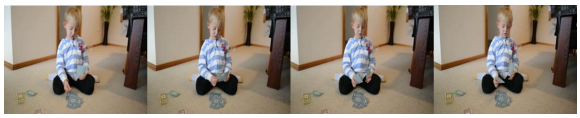
For each object $o_i \in \mathcal{R}$, we instructed the LLM to generate an alternative object $\tilde{o}_i$ belonging to the same category (e.g., replacing "car" with "truck") but with a distinct meaning, ensuring non-trivial substitutions (e.g., replacing "car" with "building"). Similarly, for each action $a_j \in \mathcal{R}$, we acquire an alternative action $\tilde{a}_j$ that the subject could perform but conveys a different meaning. For example, changing "standing" to "jumping." Note that, the LLM was instructed to generate only the plausible action that the subject can perform. Finally, the selected objects and actions in $\mathcal{R}$ are then replaced with their corresponding alternatives with the LLM, resulting in the pseudo caption $\tilde{X}$.

This process also yields a factual error explanation for $\tilde{X}$, enabling fine-grained supervision. The detailed prompts and explanation formats are provided in App. D.

Scoring. After generating a pseudo caption, assigning an intuitive quality score is essential to align with human expectations. While pretrained embeddings like visual-language model (VLM) can estimate this score, they lack a fixed scale, are constrained by context length, and do not guarantee factual grounding. Instead, we adopt a **deterministic** scoring mechanism based on factual accuracy:

$$\begin{aligned} score &= 1 - \frac{\text{\# of changed objects \& actions}}{\text{total \# of objects \& actions}} \\ &= 1 - \frac{|\mathcal{R}|}{|\mathcal{O}| + |\mathcal{A}|}. \end{aligned}$$

Since $\mathcal{R} \subseteq \mathcal{O} \cup \mathcal{A}$ and $|\mathcal{O} \cup \mathcal{A}| = |\mathcal{O}| + |\mathcal{A}|$, the score lies on a fixed scale between 0 and 1, ensuring captions with more incorrect elements receive lower scores. Since LLMs are inherently unreliable when comparing floating-point values (Spathis and Kawsar, 2023), we discretize the score into a 1-5 range ($score = round(score \times 4 + 1)$), con-



[Candidate Caption] A little girl is sitting on the floor playing with her shoes.
[Predicted Score] 3
[Explanation] The caption does not accurately capture the video content. For example, the objects (girl, shoes) are incorrect.

Figure 4: Visual example on VATEX-Eval. `VC-Inspector` produces quality assessments consistent with ground truth scores, and factual error explanations (highlighted in red). More examples are in App. A.

sistent with standard human annotation protocols (e.g., Likert-style ratings in VATEX-Eval). While this may induce minor loss of information (e.g., $0.87 \rightarrow 4$ and $0.88 \rightarrow 5$), this conversion preserves ordinal relationships and error severity while maintaining the underlying deterministic structure.

Moreover, `VC-Inspector` is jointly trained to produce textual explanations that explicitly identify incorrect objects and actions, providing fine-grained supervision beyond the scalar score.

Post-processing. We repeat the caption generation and scoring process to create 10 pseudo captions per ground truth caption, yielding 374K pseudo captions derived from 37,396 video-caption pairs. The randomized replacement of objects and actions ensures coverage across the full range of the possible scores, as the number of replacements directly influences the semantic deviation. However, this naturally results in a skewed and non-uniform score distribution. To mitigate the potential bias during training, we apply a balanced sampling strategy, resulting in a refined subset of approximately 218K pseudo captions with uniform representation across five score categories.

Due to computational constraints, training on the full 218K instances would require multiple weeks. Therefore, we further sample 44K captions (8.8K per label) from the balanced dataset for instruction tuning, which we refer to as `ActivityNet-FG-It`. This subset preserves category balance and maintains the diversity, while offering a tractable size for experimentation ($\sim$ 32 GPU hours using A100). Importantly, although we use `Llama` as the generator and perform subsampling as a practical design choice, our data generation pipeline is model-agnostic and applies across model scales and caption datasets, enabling scalable dataset construction at arbitrary sizes.

3.3 Training of `VC-Inspector`

While evaluation metrics are preferred to be lightweight, we employ the 3B/7B version of `Qwen2.5-VL` as the foundation model by finetuning it with `ActivityNet-FG-It`. To preserve the generalization capability of video features, we freeze the pretrained video encoder and the visual-language projector, and only finetune the model parameters in the LLM component with low-rank adaptation (Hu et al., 2022). Given a video-caption pair $(V, \tilde{X})$, the model predicts a quality score $S \in \{1, ..., 5\}$ along with the corresponding explanation E, i.e.,

$$[S, E] = \texttt{VC-Inspector}(V, \tilde{X}).$$

The explanation E is formatted in natural language using information collected during data generation (i.e., the list of altered objects and actions) following the template described in App. D. This can serve as extra supervision for the model to learn factual grounding and provide interpretable reasoning for this model-based evaluator at test time. Fig. 4 presents an example of the `VC-Inspector` output. During training, the model is optimized with the standard language modeling loss (Liu et al., 2023a).

4 Experiments

4.1 Experimental Settings

Training. We train `VC-Inspector` with the proposed instruction tuning dataset, `ActivityNet-FG-It`, which comprises 44K video-caption pairs, along with their quality scores and explanations.

Evaluation. The experiments evaluate the consistency and reliability of `VC-Inspector` (Section 4.3), its correlation to human judgments (Section 4.4), its sensitivity to the targeted factual errors (Section 4.5), and the contribution of individual components (Section 4.6 and 4.7). Following prior works on evaluation metrics (Shi et al., 2021; Tong et al., 2024), we report Kendall's correlation (τ_b) and Spearman's rank correlation (ρ) with ground truth scores in our main experiments.

Implementation details. Each video is uniformly sampled into 32 frames and resized to a resolution of 224×224 for both training and testing. For any image dataset, we treat them as one-frame videos. `VC-Inspector` is developed for two model sizes, 3B and 7B, initialized from their corresponding `Qwen2.5-VL` pretrained weights. In all

Metric	*ActivityNet-FG-Eval*		*YouCook2-FG-Eval*	
	τ_b	ρ	τ_b	ρ
EMScore (Shi et al., 2021)	28.94	40.77	20.21	29.24
CLIPScore (Hessel et al., 2021)	28.10	39.65	18.00	26.14
`Qwen2.5-VL-3B` (Bai et al., 2025)	37.91	47.80	37.16	47.17
`VC-Inspector-3B`	**49.53**	**62.01**	**44.29**	**55.31**

Table 1: Correlation scores on the synthetic `ActivityNet-FG-Eval` (Heilbron et al., 2015) and `YouCook2-FG-Eval` (Zhou et al., 2018) datasets. The best score is **bolded**.

experiments, we train the model on 4 NVIDIA-A100 GPUs with a global batch size of 128 and a learning rate of 1e-4. We set both alpha and rank to 32 for the low-rank adaptation with a dropout rate of 0.05. During inference, we use a temperature of 0.0 for reproducibility. The training hyperparameters are provided in App. C.

4.2 Baselines

Baselines are organized into three categories: **i) Language-based metrics** solely rely on text reference without considering the visual input. Representative metrics are BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), BERTScore (Zhang et al., 2020), SPICE (Anderson et al., 2016), SPICE-Factual (Li et al., 2023), Soft-SPICE (Li et al., 2023) and CLAIR (Chan et al., 2023). **ii) Image-augmented metrics** incorporate images as references (alongside reference captions), for better capturing the semantic alignment between the visual input and the candidate caption, e.g., CLIPScore (Hessel et al., 2021), EMScore (Shi et al., 2021), PAC-S (Sarto et al., 2023), FactVC (Liu and Wan, 2023), FLEUR (Lee et al., 2024), and G-VEVAL (Tong et al., 2024). They usually support both *reference-free* and *reference-based* settings. We report results for both settings for completeness, but our primary focus is on the *reference-free* setting, which is more practical for real-world, in-the-wild videos. **iii) Video-based metrics** employ a video encoder to incorporate full video sequences. To the best of our knowledge, no existing metric in the literature falls into this category. Therefore, we adapt CLIPScore (Hessel et al., 2021) to the recent advent of ViCLIP (Wang et al., 2024) as a stronger baseline, ViCLIPScore. Additionally, we compare `VC-Inspector` against the base model it builds upon, the vanilla `Qwen2.5-VL` model, to highlight the benefit of fine-tuning for evaluative reasoning.

Metric	*No Reference*		*1-Reference*		*9-References*	
	τ_b	ρ	τ_b	ρ	τ_b	ρ
Language-based						
BLEU_1 (Papineni et al., 2002)	-	-	12.65	16.52	28.70	36.88
BLEU_4 (Papineni et al., 2002)	-	-	12.44	14.81	22.76	25.60
ROUGE-L (Lin, 2004)	-	-	12.94	16.89	23.94	31.06
METEOR (Banerjee and Lavie, 2005)	-	-	16.68	21.80	27.64	35.76
CIDEr (Vedantam et al., 2015)	-	-	17.62	23.02	27.92	36.18
BERTScore (Zhang et al., 2020)	-	-	15.24	19.82	25.05	32.37
SPICE (Anderson et al., 2016)	-	-	14.80	18.78	27.41	35.40
SPICE-factual (Li et al., 2023)	-	-	13.59	17.04	26.05	35.58
Soft-SPICE (Li et al., 2023)	-	-	21.25	27.61	36.31	46.41
CLAIR* (Chan et al., 2023)	-	-	36.00	-	34.80	-
Multimodal - image-based						
CLIPScore (Hessel et al., 2021)	22.33	29.09	27.39	35.49	35.21	45.28
EMScore (Shi et al., 2021)	22.88	29.79	28.63	37.05	36.66	**47.00**
FactVC (Liu and Wan, 2023)	22.79	29.69	28.78	**37.22**	36.18	46.33
PAC-S* (Sarto et al., 2023)	25.10	-	31.40	-	38.10	-
G-VEval* (Tong et al., 2024)	39.40	-	**44.90**	-	**48.10**	-
Multimodal - video-based						
ViCLIPScore (Wang et al., 2024)	30.92	39.86	-	-	-	-
`Qwen2.5-VL-3B` (Bai et al., 2025)	31.29	36.43	-	-	-	-
`Qwen2.5-VL-7B` (Bai et al., 2025)	34.70	39.40	-	-	-	-
`VC-Inspector-3B`	37.99	42.45	-	-	-	-
`VC-Inspector-7B`	**42.58**	**45.99**	-	-	-	-

Table 2: Human correlation scores on the VATEX-Eval (Shi et al., 2021) dataset. * indicates results reported from the original paper. The best of each column section is **bolded**. Please note that this work focuses on the *No Reference* setting. **Our best model outperforms all other models in this setting, while remaining competitive with reference-based metrics.**

4.3 Consistency and Reliability of Quality Estimates

We first examine whether the quality estimation of `VC-Inspector` is consistent across visual domains. Following the same data generation pipeline in Section 3.2, we create two evaluation datasets, `ActivityNet-FG-Eval` and `YouCook2-FG-Eval`, according to the ActivityNet (Krishna et al., 2017) test set and YouCook2 (Zhou et al., 2018) validation set, respectively. Both datasets contain pseudo captions with varying degrees of factual errors.

Table 1 presents the correlation between the metric scores and ground truth annotations in these two datasets. `VC-Inspector`, specifically finetuned for factual grounding, consistently outperforms baselines in differentiating incorrect captions. Notably, although it is only trained on `ActivityNet-FG-It`, its quality estimation generalizes effectively across visual domains. This suggests that the proposed data generation pipeline produces stable and reliable quality estimates, rather than noisy outputs.

4.4 Correlation with Human Judgments

Evaluation on VATEX-Eval. Table 2 presents our primary results: human correlation scores on the VATEX-Eval (Shi et al., 2021) dataset. VATEX-Eval is a widely adopted dataset for evaluating

Metric	τ_b	ρ
PAC-S (ViT-B) (Sarto et al., 2023)	27.8	34.3
PAC-S (ViT-L) (Sarto et al., 2023)	42.9	54.8
EMScore (Shi et al., 2021)	58.9	73.9
VC-Inspector-7B	**72.8**	**80.3**

Table 3: Human correlation scores on the YouCook2-Eval dataset.

video caption metrics. It contains 6 captions with varying levels of quality per video, each rated by three human evaluators on a scale of 1 to 5.[1]

Baselines are grouped by reference type as outlined in Section 4.2, and the three column-sections correspond to the *No Reference*, *1-Reference*, and *9-References* settings, respectively. Language-based metrics, which rely heavily on textual references, are not applicable to our target, *No Reference* setting. Therefore, we focus our comparisons on multimodal methods that incorporate visual inputs.

`VC-Inspector` consistently outperforms all evaluated metrics in the *reference-free* setting, particularly those based on (image-based) CLIP embeddings. When adapting CLIPScore to the recently introduced ViCLIP (Wang et al., 2024) (which adopts a video encoder), we observe a considerable gain over those image-CLIP-based approaches. Despite ViCLIPScore being a stronger baseline, it still underperforms relative to `VC-Inspector` and its underlying model, as the context length of the ViCLIP model (i.e., 32) is much shorter than the 32K context length of ours, which supports a more flexible and comprehensive caption evaluation. Furthermore, `VC-Inspector` outperforms G-VEval (Tong et al., 2024) despite being based on GPT-4o, a substantially larger proprietary model. This demonstrates that explicit factual grounding and explanation supervision are more critical than model scale alone. In contrast, `VC-Inspector` is open-source, lightweight, and reproducible, with configurations available at 3B and 7B parameters depending on system requirements. While having a more compact size relative to G-VEval, our 7B model achieves the highest correlation to human evaluations and could potentially enable on-the-fly quality estimation during training, making it a viable reward model in Reinforcement Learning (RL) applications. Although we focus on the *reference-free* setting, it is noteworthy that `VC-Inspector` even surpasses the performance of most *reference-based* metrics.

[1]Since some videos become unavailable, we collect the remaining subset of 2,590 videos and the corresponding 15,540 candidate captions. Unless otherwise specified, all experiments were evaluated on the same dataset to ensure fairness.

Metric	*Flickr8K-Expert*	*Flickr8K-CF*
Reference-based		
BLEU_1 (Papineni et al., 2002)	32.2	17.9
BLEU_4 (Papineni et al., 2002)	30.6	16.9
ROUGE (Lin, 2004)	32.1	19.9
METEOR (Banerjee and Lavie, 2005)	41.5	22.2
CIDEr (Vedantam et al., 2015)	43.6	24.6
SPICE (Anderson et al., 2016)	51.7	24.4
BERTScore (Zhang et al., 2020)	-	22.8
CLIPScore (Hessel et al., 2021)	52.6	36.4
PAC-S (Sarto et al., 2023)	55.5	37.6
FLEUR (Lee et al., 2024)	51.6	<u>38.8</u>
HICE-S (Zeng et al., 2024)	<u>57.2</u>	38.2
PAC-S++ (ViT-B) (Sarto et al., 2025)	55.3	37.9
PAC-S++ (ViT-L) (Sarto et al., 2025)	-	<u>38.8</u>
Polos (Wada et al., 2024)	56.4	37.8
Reference-free		
CLIPScore (Hessel et al., 2021)	51.1	34.4
PAC-S (Sarto et al., 2023)	53.9	36.0
FLEUR (Lee et al., 2024)	52.7	38.6
HICE-S (Zeng et al., 2024)	55.9	37.2
PAC-S++ (ViT-B) (Sarto et al., 2025)	54.1	37.0
PAC-S++ (ViT-L) (Sarto et al., 2025)	-	38.5
`VC-Inspector-3B`	59.9	39.0
`VC-Inspector-7B`	**<u>63.4</u>**	**<u>46.0</u>**

Table 4: Correlation score (τ_b) with human judgments on Flickr8k-Expert and Flickr8k-CF (Hodosh et al., 2013) dataset. The overall best scores are **bolded** and best of each section is <u>underlined</u>. **Our model outperforms even the reference-based methods.**

Evaluation on YouCook2-Eval. Due to the limited availability of public benchmarks for video caption metric evaluation, we construct another video caption dataset with human ratings, `YouCook2-Eval`, based on YouCook2 (Zhou et al., 2018) validation set. We randomly sample 20 videos, each paired with 5 candidate captions including the ground truth caption, a caption from another video, and captions generated from three different models (i.e., Gemini2.5 (Comanici et al., 2025), Qwen2.5-VL (Bai et al., 2025), PDVC (Wang et al., 2021)), resulting in 100 video-caption pairs with diverse quality. Following VATEX-Eval (Shi et al., 2021), we collect human ratings from 3 human annotators on a 1-5 Likert scale, and evaluate the correlation to human judgments. As shown in Table 3, `VC-Inspector` achieves the highest correlation compared to baselines, showing generalization to a new domain and caption sources.

Evaluation on Flickr8K. We further extend our evaluation to two popular image caption datasets, Flickr8K-Expert and Flickr8K-CF (Hodosh et al., 2013), by viewing images as single-frame videos. The former contains 17K image-caption pairs rated

Metric	*FOIL-COCO*	*ActivityNet-FOIL*
Reference-based		
EMScore (Shi et al., 2021)	-	92.4
PAC-S (Sarto et al., 2023)	93.5	93.4
FLEUR (Lee et al., 2024)	97.3	-
Reference-free		
EMScore (Shi et al., 2021)	-	89.5
PAC-S (Sarto et al., 2023)	90.2	91.0
FLEUR (Lee et al., 2024)	96.8	-
VC-Inspector-3B	**99.6**	**99.3**

Table 5: Accuracy on image/video caption hallucination benchmarks.

by three human experts on a scale of 1 to 4, which requires a more fine-grained differentiation among the captions, whereas the latter collects binary quality assessments for 48K image-caption pairs from crowd sources. Table 4 reports the human correlation score on these two datasets. VC-Inspector is instruction-tuned with captions that exhibit varying degrees of factual inaccuracies, enabling nuanced evaluation of caption quality. This allows the model to perform effectively across both benchmarks, despite their differing rating scheme. In these evaluations, VC-Inspector remains the best-performing method under the *reference-free* setting, and even outperforming several *reference-based* metrics. These results demonstrate the strong generalization capability of VC-Inspector across visual domains and video lengths.

4.5 Sensitivity to Hallucinations

Table 5 reports accuracy on two hallucination detection benchmarks, FOIL-COCO (Shekhar et al., 2017) and ActivityNet-FOIL (Shi et al., 2021), where the task is to classify FOIL captions from their corresponding correct captions. Each FOIL caption differs from the original by exactly one object error, enabling a controlled evaluation of hallucination sensitivity. As VC-Inspector is explicitly designed to assess object and action grounding in visual-caption alignment, it performs strongly on both image-based (FOIL-COCO) and video-based (ActivityNet-FOIL) benchmarks. These results indicate that VC-Inspector is highly sensitive to targeted hallucination errors and can reliably identify factual inconsistencies across varying video lengths and visual contexts.

4.6 Ablation on Data Synthesis Strategies

Since both objects and actions are key factual elements in video captions, grounding the evaluator in factual understanding requires instructing the model to identify errors in these components. To this end, we systematically altered both elements in ground truth captions during data generation to create pseudo captions for training. In this section, we ablate the impact of modifying these elements in the ActivityNet training set by evaluating three variants: i) Changing objects only, ii) Changing actions only, and iii) Changing both. As shown in Table 6, all variants exhibit strong alignment with human ratings compared to EMScore (Shi et al., 2021). However, the variant that alters both objects and actions yields the best performance. These results underscore the importance of both factual elements, objects and actions, for capturing video context in caption evaluation. They also demonstrate the robustness and generalization of the proposed paradigm across different factual errors.

Metric	Data Synthesis	τ_b	ρ
EMScore (Shi et al., 2021)	n/a	22.88	29.79
VC-Inspector-3B	Change objects only	36.40	41.20
	Change actions only	33.23	39.63
	Change both (Ours)	**37.99**	**42.45**

Table 6: Ablation study on synthetic data generation strategies. Human correlation scores (τ_b, ρ) are reported on the VATEX-Eval benchmark.

4.7 Analysis of Explanations

Impact of explanation on performance. Our data generation process produces not only pseudo captions and quality scores but also an informative "side product": explanations identifying factual errors in candidate captions. We leverage these explanations as auxiliary supervision during training, enabling the model to learn better factual grounding, which has been shown to be effective compared to the variant without explanations in Table 7. Beyond training benefits, explanations enhance the interpretability of the model-based metric, providing transparency of why specific scores are assigned.

Using explanations for caption refinement. Having shown that explanations aids in evaluating caption quality, we now ask: **can they also help improve the captions themselves?** To assess this, we adopt Qwen2.5-VL-7B as the captioning model and use VC-Inspector-7B as the critique model that provides feedback during an iterative refinement process, repeated for up to 10 iterations. The prompt used for refinement is provided in D.8. We conduct this experiment on CAPability (Liu et al., 2025), a recent benchmark that evaluates

VC-Inspector-3B	τ_b	ρ
Without Explanations	34.29	38.18
With Explanations	**37.99**	**42.45**

Table 7: Impact of explanations on the model performance on VATEX-Eval. In the "Without Explanation" setting, we have trained the model only on the score, removing the pseudo explanations.

	Feedback Provider		
Task	None	Qwen2.5-VL-7B	VC-Inspector-7B
Dynamic Object Num.	17.4	17.3	**20.7**
Event	58.2	18.8	**58.6**
Object Category	67.8	58.9	**68.0**
Object Number	28.8	20.4	**29.3**
Spatial Relation	56.6	45.7	**57.1**

Table 8: F1 scores across different tasks. Video-based tasks are highlighted in blue, while image-based tasks are highlighted in beige. Best results are shown in **bold**.

caption quality across 12 distinct dimensions and two modalities (image and video), with human-annotated ground truth. Following its protocol, final captions are scored by GPT-4.1. Note that we report dimensions related to object semantics and events in Table 8, and provide the full results in App. B.2. We compare against baselines using the same captioner but with different feedback strategies: (a) no feedback, and (b) feedback from the off-the-shelf Qwen2.5-VL-7B. Results show that the captioner guided by VC-Inspector consistently improves over both baselines on these dimensions, demonstrating the effectiveness of explanations for caption refinement. Notably, comparing the performance of (b) to (a), we find that undesired feedback to the captioner can even hurt the performance, significantly. This further underscores the importance of explicit factual grounding in VC-Inspector. App. B.2 additionally presents a qualitative example of the refinement, showing that VC-Inspector successfully guides the correction of factual errors in the initial caption.

Evaluation of explanation quality. For a more direct assessment of explanation quality, we randomly sample 50 captions from the VATEX-Eval dataset, and manually evaluate the explanations generated by VC-Inspector-7B. For each explanation, two human evaluators are provided with the corresponding video and candidate caption, and rate the explanation on a 1-5 Likert scale rather than binary judgments (e.g., good vs. bad), as explanations may be correct in some aspects while missing others. To mitigate the potential bias toward assigning high ratings, we introduced a control condition where, for half of evaluation data, the explanation was randomly paired with a different video caption. The scores are normalized to $[0, 1]$, and the final score is computed as the score on correct pairs, discounted by the score assigned to incorrect pairs, i.e., $final\_score = (score\ on\ correct\ pairs) \times (1 - score\ on\ incorrect\ pairs)$. VC-Inspector achieves a final score of 0.62, with 86% inter-evaluator agreement within one point. In addition to human evaluation, we report language model based evaluation in App. B.1.

4.8 Computational Efficiency

We assess computational efficiency by comparing the average runtime per video. On a single A100 GPU, EMScore (Shi et al., 2021), ViCLIPScore (Wang et al., 2024), and VC-Inspector-3B require 0.42, 0.34, and 0.30 seconds per clip, respectively. These results indicate that VC-Inspector is more efficient than existing methods. For EMScore and ViCLIPScore, we use the official implementations provided by the authors; for VC-Inspector, we employ vLLM (Kwon et al., 2023) with online serving. Efficiency is measured over 2,000 videos, each uniformly sampled into 16 frames.

5 Conclusion

This work tackles the challenge of evaluating video captions across diverse domains without relying on human-annotated references. We conducted an extensive review of existing metrics, identified their limitations, and proposed VC-Inspector, a novel *fact-aware* evaluator for *reference-free* video caption evaluation. By incorporating factual analysis into caption assessment, we equipped a lightweight, open-source LMM with the factual grounding capabilities through training on captions of diverse quality, where pseudo captions with controllable factual errors were systematically generated. Experimental results across multiple domains demonstrate that VC-Inspector achieves high alignment with human judgments, and outperforms existing metrics in detecting factual errors. Its versatility and interpretability make it a practical tool for assessing factual accuracy in real-world video captioning scenarios. Furthermore, ablation studies on predicted explanations underscore its effectiveness in quality estimation and caption refinement.

Limitations

VC-Inspector primarily targets object and action correctness, which constitute major sources of factual errors in video captions. While these factors cover many common failure cases, other important aspects, e.g., attributes, spatial relationships, and fine-grained temporal ordering, are not explicitly modeled and remain directions for future work. Additionally, the training process partially relies on synthetically generated captions and pseudo-scores. Although these scores are deterministic and validated through correlation with human judgments, they may not fully capture the diversity of real-world captioning errors. Finally, while VC-Inspector is substantially more lightweight than proprietary LLM-based evaluators, it still depends on a multimodal backbone, resulting in higher computational costs compared to purely embedding-based metrics.

Despite these limitations, VC-Inspector *represents a meaningful step toward* ***explanation-aware***, ***factually grounded***, *and* ***reference-free*** *video caption evaluation, providing a flexible foundation for future extensions.*

Acknowledgment

Some experiments were conducted on the UMBC HPCF, supported by the National Science Foundation under Grant No. CNS-1920079.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report.

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

David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. 2023. CLAIR: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13638–13646, Singapore. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, abs/2507.06261.

Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. 2018. Learning to evaluate image captioning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5804–5812. IEEE Computer Society.

Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. 2020. Multi-modal transformer for video retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 214–229. Springer.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 961–970. IEEE Computer Society.

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

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152, Hong Kong, China. Association for Computational Linguistics.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 706–715. IEEE Computer Society.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226, Online. Association for Computational Linguistics.

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

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An Explainable Reference-Free Evaluation Metric for Image Captioning Using a Large Multimodal Model.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023. FACTUAL: A benchmark for faithful and consistent textual scene graph parsing. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6377–6390, Toronto, Canada. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hui Liu and Xiaojun Wan. 2023. Models see hallucinations: Evaluating the factuality in video captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11807–11823, Singapore. Association for Computational Linguistics.

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

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Pandeng Li, Boqiang Zhang, Nianzu Yang, Zuan Gao, Yun Zheng, Hongtao Xie, and 1 others. 2025. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2019. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, Florence, Italy. Association for Computational Linguistics.

Koki Maeda, Shuhei Kurita, Taiki Miyanishi, and Naoaki Okazaki. 2024. Vision Language Model-based Caption Evaluation Method Leveraging Visual Context Extraction.

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

Jeshmol P.J. and Binsu C. Kovoor. 2024. Video question answering: A survey of the state-of-the-art. *Journal of Visual Communication and Image Representation*, 105:104320.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. 2024. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24*

July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 6914–6924. IEEE.

Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Positive-augmented contrastive learning for vision-and-language evaluation and training. *International Journal of Computer Vision*, 133(11):7647–7671.

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

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2021. EMScore: Evaluating Video Captioning via Coarse-Grained and Fine-Grained Embedding Matching.

Dimitris Spathis and Fahim Kawsar. 2023. The first step is the hardest: Pitfalls of Representing and Tokenizing Temporal Data for Large Language Models.

Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. 2024. G-VEval: A Versatile Metric for Evaluating Image and Video Captions Using GPT-4o. volume abs/2412.13647.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Yuiga Wada, Kanta Kaneda, Daichi Saito, and Komei Sugiura. 2024. Polos: Multimodal metric learning from human feedback for image captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13559–13568. IEEE.

Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. 2021. End-to-end dense video captioning with parallel decoding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6827–6837. IEEE.

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. 2024. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yanzhi Yi, Hangyu Deng, and Jinglu Hu. 2020. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 985–994, Online. Association for Computational Linguistics.

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. volume abs/2407.18589.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

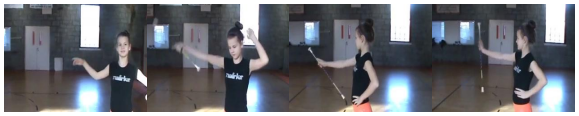
Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598. AAAI Press.

Appendix

A Qualitative Results

Fig. 5 presents visual examples of `VC-Inspector` outputs on the `ActivityNet-FG-Eval` (top) and VATEX-Eval (others) datasets. Without relying on reference captions, the model evaluates candidate captions based on the associated video content and produces quality scores that closely mirror human judgments. The explanation effectively pinpoints the factual inaccuracies, such as incorrect objects and/or actions in the candidate captions, and assigns scores accordingly.

In Fig. 6, we showcase two video examples from `ActivityNet-FG-Eval`, each paired with candidate captions containing a progressively increasing number of factual errors. The model successfully detects these incorrect elements, provides detailed explanations, and yields scores that reflect the severity of the factual inaccuracies.



[Candidate Caption] Then she does several dance moves with the sword.
[Predicted Score] 4
[Explanation] The caption does not accurately capture the video content. For example, the objects (sword) are incorrect.

[Candidate Caption] A man screws a screw gun into a thin strip of wood.
[Predicted Score] 3
[Explanation] The caption does not accurately capture the video content. For example, the objects (screw gun) and actions (screws) are incorrect.

[Candidate Caption] A young man does the long jump at a track and field event.
[Predicted Score] 5
[Explanation] The caption is helpful, relevant, accurate, and informative to the video content.

Figure 5: Visual examples from `ActivityNet-FG-Eval` (top) and VATEX-Eval (others). `VC-Inspector` produces quality assessments consistent with ground truth scores, and explanatory insights into factual errors (highlighted in red).

VC-Inspector

[Candidate Caption] A little girl is sitting on a chair.
[Predicted Score] 5
[Explanation] The caption is helpful, relevant, accurate, and informative to the video content.

[Candidate Caption] A little girl is sleeping on a chair.
[Predicted Score] 4
[Explanation] The caption does not accurately capture the video content. For example, the actions (sleeping) are incorrect.

[Candidate Caption] A woman is sleeping on a chair.
[Predicted Score] 2
[Explanation] The caption does not accurately capture the video content. For example, the objects (woman) and actions (sleeping) are incorrect.

VC-Inspector

[Candidate Caption] Three people are dancing in the room.
[Predicted Score] 5
[Explanation] The caption is helpful, relevant, accurate, and informative to the video content.

[Candidate Caption] Three people are chatting in the room.
[Predicted Score] 4
[Explanation] The caption does not accurately capture the video content. For example, the actions (chatting) are incorrect.

[Candidate Caption] Three people are chatting in the garden.
[Predicted Score] 2
[Explanation] The caption does not accurately capture the video content. For example, the objects (garden) and actions (chatting) are incorrect.

Figure 6: Additional visualization of `VC-Inspector` results on `ActivityNet-FG-Eval` videos, with candidate captions of diverse quality. Incorrect objects and actions are identified by `VC-Inspector` and labeled in red.

B Additional Results

B.1 Evaluation of Explanation Quality using Language Models

We further assess the quality of the explanations generated by `VC-Inspector` on two synthetic datasets, as the VATEX-Eval dataset does not offer ground truth explanations. Table 9 presents two metrics for this experiment. The BERT score measures the semantic similarity between the generated explanations and the pseudo ground truth explanations, according to pretrained textual embeddings. On the other hand, the LLM Score is derived by using the prompt detailed in Section D to engage `Llama-3-8B-Instruct` in a LLM-as-a-judge approach, similar to (Liu et al., 2023a). We report the relative score assigned by the LLM that rescales the score received by the predicted explanations with respect to the score given to the ground truth explanations. Results show that the explanations provided by `VC-Inspector`, along with its quality assessment, align closely with the (pseudo) ground truth explanations obtained during the data generation process.

Dataset	BERT Score	LLM Score
ActivityNet-FG-Eval	0.79	93.11
YouCook2-FG-Eval	0.70	90.97

Table 9: Evaluation of the explanations generated by `VC-Inspector-3B` on two synthetic evaluation datasets.

B.2 Complete Results for Caption Refinement

In Section 4.7, we present the experiment of caption refinement with CAPability benchmark, where we report the dimensions concerning objects and actions, as they are the primary focus of this work. We provide the F1 scores of all the 12 dimensions in Table 10 for completeness. In addition to quantitative results, Fig. 7 visualizes an example of the refinement, showing the initial and refined captions produced under guidance from `VC-Inspector`.

C Training Hyperparameters

We report all training hyperparameters in Table 11.

Task	Feedback Provider		
	None	Qwen2.5-VL-7B	VC-Inspector-7B
Camera Angle	**37.5**	6.4	36.1
Camera Movement	26.4	0.6	**29.3**
Character Identification	29.6	17.3	**30.5**
Dynamic Object Num.	17.4	17.3	**20.7**
Event	58.2	18.8	**58.6**
Object Category	67.8	58.9	**68.0**
Object Color	**73.0**	51.9	72.3
Object Number	28.8	20.4	**29.3**
OCR	**69.6**	62.9	68.5
Scene	**63.5**	63.0	63.2
Spatial Relation	56.6	45.7	**57.1**
Style	**84.3**	78.7	84.2
Average	51.1	36.8	**51.5**

Table 10: F1 scores across different tasks. Video-based tasks are highlighted in blue, while image-based tasks are highlighted in beige. Best results are shown in **bold**.

Hyperparameter	Value
Training Configuration	
Epochs	1
Global Batch Size	128
Learning Rate	1e-4
LR Scheduler	Cosine (min: 1e-5)
Warmup Ratio	0.05
LoRA Configuration	
Rank (r)	32
Alpha (α)	32
Dropout	0.05
Module Frozen	
Vision Encoder	✓
Language Model	✗
Projector	✓
Video Processing	
Number of Frames	32
Max. Pixels	224×224

Table 11: Hyperparameters for instruction fine-tuning.

D Prompts

The data generation prompts are reported on the following blocks:

- Extract object – Prompt D.1
- Extract action – Prompt D.2
- Find similar object – Prompt D.3
- Find similar action – Prompt D.4
- Substitute object or action – Prompt D.5

The data generation process provides the information of incorrect objects and actions. We format the explanations of these factual errors using the following template:

- Captions without errors: The caption is helpful, relevant, accurate, and informative to the video content.
- Captions with errors: The caption does not accurately capture the video content. For example, the actions ({{wrong_act}}) are incorrect / the objects ({{wrong_obj}}) are incorrect / the objects ({{wrong_obj}}) and actions ({{wrong_act}}) are incorrect.

These explanations are then used in Prompt D.6 for training VC-Inspector.

The prompt for evaluating the generated explanations is in Prompt D.7. Prompt D.8 is used to refine the caption based on the explanation.

Prompt D.1: Extract object from a caption

```
### Instruction:

Given the input text, generate a list of
objects in the caption in the format of [
"Object1", "Object2", ...]. Don't include any
verbs. ONLY REPLY THE ANSWER.

### Input: {{caption}}
### Output:
```

Prompt D.2: Extract actions from a caption

```
### Instruction:

Given the input text, generate a list of
actions in the caption in the format of
["Action1", "Action2", ...]. ONLY REPLY THE
ANSWER.

### Input: {{caption}}

### Output:
```

Prompt D.3: Find similar object given an object

```
### Instruction:

Find the parent class of the given object and
generate one of its child classes that has a
different meaning but shares the same parent.
The new class cannot be a synonym or similar
term to the original object. It can be an
antonym or any co-hyponym. For example,
generate "dog" for "cat". ONLY REPLY THE NEW
CLASS.

### Input: {{object}}

### Output:
```

Prompt D.4: Prompt to find similar action given an action

```
### Instruction:
```

```
Find a different action that the subject can
perform that has a different meaning than the
input action. The new action cannot be a
synonym or similar term to the original
action. For example, generate "put into" for
"take out of". ONLY REPLY THE NEW ACTION.

### Input: {{action}}

### Output:
```

Prompt D.5: Prompt to substitute object or action given the caption and new object or actions

```
### Instruction:

Substitute {{old_obj_act}} in {{caption}} as
{{new_obj_act}}. Keep the answer in the same
format as {{caption}}. ONLY REPLY THE ANSWER.

### Input: {{caption}}

### Output:
```

Prompt D.6: Fine-tuning prompt

```
### USER:
{{VIDEO}}

<caption>{{caption}}</caption> You are given
a video and a caption describing the video
content. Please rate the helpfulness,
relevance, accuracy, level of detail of the
caption. The overall score should be on a
scale of 1 to 5, where a higher score
indicates better overall performance. Please
first output a single line containing only
one integer indicating the score. In the
subsequent line, please provide a
comprehensive explanation of your evaluation,
avoiding any potential bias. STRICTLY FOLLOW
THE FORMAT.

### ASSISTANT:
{{quality_score}}

{{explanation}}
```

Prompt D.7: Prompt for explanation evaluation

```
[Context]
{{ground_truth_caption}}

[Caption]
{{caption_to_evaluate}}

[Groundtruth]
{{ground_truth_explanation}}
[End of Groundtruth]

[Assistant]
{{predicted_explanation}}
[End of Assistant]

[System]
We would like to request your feedback on the
performance of an AI assistant in the
response to the quality evaluation of the
caption provided above with respect to a
video. For your reference, the visual content
in the video is represented with a few
sentences describing the same video. You are
also given a ground truth evaluation to that
caption.
Please rate the helpfulness, relevance,
accuracy, level of detail of the response by
comparing to the ground truth and referring
to the context information. Provide an
overall score on a scale of 1 to 10, where a
higher score indicates better overall
performance.

Please first output a single line containing
the score. In the subsequent line, please
provide a comprehensive explanation of your
evaluation, avoiding any potential bias.
```

Prompt D.8: Prompt to refine caption based on explanation

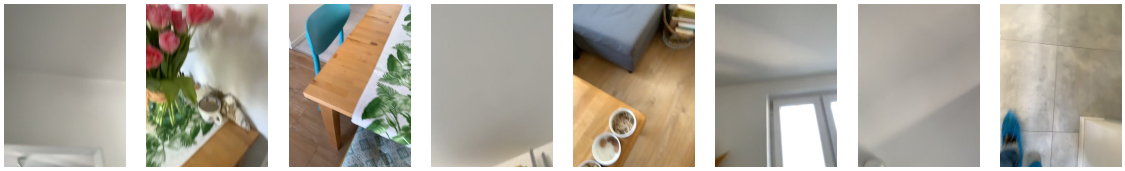
```
Previous Caption: {{caption}}
Evaluation: {{explanation}}

Based on the previous caption and evaluation,
refine the caption to make it more accurate
and detailed.
Use the evaluation to find out the missing
details and use the video frames to fix the
missing details.
Just output the refined caption, without any
other text.
```

E Use of AI Assistance

During the preparation of this work, the authors utilized Cursor[2] for coding assistance and ChatGPT[3] for proofreading and language refinement. All input provided to these tools consisted of original content authored by the authors.

[2] https://cursor.com
[3] https://chatgpt.com/



t=2.7 s | t=8.2 s | t=13.6 s | t=19.1 s | t=24.5 s | t=30.0 s | t=35.5 s | t=40.9 s

Round 0
3 / 5

The video begins with a clear, focused shot of a vintage-style poster for the "Chicago World's Fair" from 1933-1934, which is laid flat on a light gray tiled floor. The poster features a striking red background with an illustration of the fair's iconic buildings and the text "CHICAGO WORLD'S FAIR MAY 27 - NOV. 1, 1933" at the top.
The camera then moves erratically, capturing blurry glimpses of the surrounding environment. It pans across a hallway or room corner, showing a white wall and a glimpse of a wooden door frame. The view shifts to a dining area where a wooden table is covered with a green patterned tablecloth, accompanied by a blue chair. A vase with pink flowers sits on the table, and a potted plant is visible nearby.
The camera continues its shaky movement, briefly focusing on a yellow chair with a patterned seat cushion, before moving to show a dark-colored sofa with pillows against a wall. A small side table with books is next to the sofa.
The camera then pans upward, showing the ceiling and a light fixture.
The perspective changes again, revealing a yellow ottoman near a wooden cabinet or shelving unit. The camera then moves back toward the hallway, showing the tiled floor and a glimpse of a doorway leading to another room.
The final frames return to the poster, now viewed from a slightly different angle, still lying on the tiled floor, with a person's blue shoe visible in the foreground. The video ends with a clear, steady shot of the poster, emphasizing its design and historical significance.

Round 5
5 / 5

The video opens with a clear, steady shot of a vintage-style poster for the "Chicago World's Fair" from 1933-1934, laid flat on a light gray tiled floor. The poster features a striking red background with an illustration of the fair's iconic buildings and the text "CHICAGO WORLD'S FAIR MAY 27 - NOV. 1, 1933" at the top.
The camera then moves erratically, panning and tilting rapidly through various parts of a modern, well-lit apartment interior. Due to the quick, unsteady movements, many shots are blurry, but glimpses of different areas and objects can be discerned:

- A hallway with a white wall and a wooden door frame.
- A dining area featuring a wooden table, a bright blue chair, and a vase of pink flowers.
- A living room with a dark gray sofa, a yellow ottoman, and a wooden shelf mounted on the wall holding books and decorative items.
- A view of a kitchen area with white cabinets and a glimpse of a refrigerator.
- A close-up of a person's blue shoe on the floor next to the poster.
- The poster is shown again, this time from a slightly different angle, still lying on the tiled floor.

The video concludes with a final clear shot of the Chicago World's Fair poster, emphasizing its design and historical significance against the backdrop of the contemporary interior. The overall impression is that of a casual, perhaps hurried, tour of a stylishly furnished home, with the poster serving as a point of interest or a decorative element.

Figure 7: Qualitative example from the DYNAMIC OBJECT NUMBER split of CAPability captioned by `Qwen2.5-VL-7B`. **Round 0** is the raw caption with no self-refinement. After five rounds of iterative refinement, **Round 5** restructures the shaky mid-section into an explicit room-by-room inventory, and the CAPability score improves from 3/5 to 5/5. Red shading marks errors in Round 0 that Round 5 corrects: the yellow ottoman is misidentified as a chair, and the wicker basket of books is described as a side table. Green shading marks the corresponding Round 5 fixes, together with the newly-covered kitchen area that Round 0 omitted.