

Evaluating Vision-Language Models Against Automatic Metrics for Video Captioning

Navya Rachakonda^a and George Awad^a

^aNational Institute of Standards and Technology, Gaithersburg, MD, USA 20899

ABSTRACT

The task of captioning videos has several applications, such as in defense and security video analysis, human-robot interaction, a visual tool for the hearing/visually impaired, video summarization, automatic scene description, and supporting video indexing.^{1,2} Traditional video caption evaluation has been adopting text-based metrics such as BLEU, METEOR, and CIDER.³ However, one key limiting feature is that they often fail to capture the semantic alignment with the visual content, that is, their output does not always align with the image content. Consequently, this limitation was addressed by exploring whether vision-language models (VLMs), in particular CLIP, ALIGN, and LLAMA 3.2 11B Vision(LLAMA),⁴ can serve as alternative evaluation tools. In this work, we propose a reference-free evaluation approach that extracts keyframes from videos, computes similarity scores between keyframes and captions using VLMs, and analyzes their correlation with traditional metrics. We used a benchmark dataset of 1,757 videos and 31,626 captioning systems' outputs. We observe weak-to-moderate correlations between VLM-based similarity scores and traditional evaluation metrics. Our findings suggest that while VLMs can capture aspects of semantic alignment that traditional metrics cannot capture, they are still not able to fully replace conventional approaches just yet.

Keywords: Vision Language Models, VLMs, Captioning systems

1. INTRODUCTION

Evaluating the quality of automatic video captioning models requires large-scale datasets, human annotations, and robust metrics. Evaluation campaigns such as TRECVID⁵ have been evaluating video-to-text (VTT) tasks for several years, using both human annotations and automatic metrics. Annotating videos by humans is a time-consuming and expensive process. There has also been limited research into whether VLMs can provide reliable reference-free evaluation based on semantic image-caption alignment. To address this, we explored VLMs, which integrate visual and textual information to perform tasks. They can take a variety of input modalities, such as text, images, videos, and generate inferences on those inputs. Traditional evaluation metrics are based on text comparison, where system-generated captions are compared against reference captions using n-gram overlap or lexical similarity.⁶ In contrast, our method performs an image-to-text evaluation using VLMs to process both the image and the caption. Our goal is to understand (1) whether VLM-based similarity scores correlate with scores generated by traditional automatic metrics and (2) whether VLMs can replace human-generated annotations. Our contribution in this paper can be summarized as follows: (1) a novel reference-free evaluation framework using VLMs, (2) a large-scale study on 1,757 videos and 31,626 captions, and (3) a comprehensive correlation analysis between VLM-based similarity scorers and traditional evaluation metrics.

The rest of the paper is organized as follows: In section 2, we review the state-of-the-art of video captioning metrics, section 3 discusses the data and experimental design followed by discussion of results in section 4, and finally we conclude in section 5.

2. RELATED WORKS

In video captioning research, most automatic evaluation metrics are borrowed from the field of Natural Language Processing (NLP).⁷ They have been widely used to assess the performance of video captioning systems, as they became popular since the launch of the Large Scale Movie Description Challenge (LSMDC) in 2015.⁸ The methods and metrics in this work can be broadly categorized as reference-based metrics and reference-free metrics, and they are described next.

2.1 Reference-based metrics:

The Bilingual Evaluation Understudy (BLEU)⁹ is an automatic evaluation metric that is used in machine translation and captioning to measure how closely a generated sentence matches human-written reference sentences. It is based on comparing words and phrases that overlap (called n-grams) between the two sentences, using a modified precision of the n-grams, which is the proportion of the matched n-grams out of the total number of n-grams in the evaluated translation¹⁰ and a brevity penalty factor to penalize candidate sentences shorter than reference sentences compared to candidate sentences. BLEU scores range from 0 to 1, with higher scores indicating greater similarity between the candidate and the reference sentences. However, BLEU mainly focuses on word overlap rather than on observing true semantic meaning. This may lead BLEU to fail to fully capture how well a sentence conveys meaning.

The Metric for Evaluation of Translation with Explicit ORdering (METEOR)¹¹ was proposed to address the shortcomings of BLEU. This includes a lack of recall, the use of higher-order n-grams, lack of explicit word-matching between translation and reference, and the use of geometric averaging n-grams. Using the reference sentence and the generated sentence, METEOR creates an alignment between the unigrams in the two sentences, where each unigram maps to 0 or 1 unigram in the other sentence, and to no unigrams in the same sentence. This alignment is made incrementally through a series of stages. And unlike BLEU, METEOR considers both precision and recall, as well as stemmed words, synonyms, and word order to better capture semantic meaning. METEOR often shows a stronger correlation with human judgment compared to BLEU.

The Consensus-based Image Description Evaluation (CIDEr)¹² is another automatic evaluation metric specifically designed to evaluate image captioning approaches. Compared to traditional n-gram metrics, CIDEr applies TF-IDF (term frequency-inverse document frequency) weighting to n-grams. This is used to emphasize informative and image-specific words while reducing the influence of repeated phrases. This metric computes sentence similarity using cosine similarity between the reference and generated sentences across unigram to 4-gram levels. CIDEr has been shown to correlate more strongly with human judgment of semantic relevance than metrics such as BLEU.

The CIDEr-D¹² metric is a modification of CIDEr, used to evaluate image and video captioning approaches. It was made to help defend against the gaming of the CIDEr metric, as CIDEr might produce sentences with high scores but low results when judged by a human. Some of the modifications include the removal of stemming, which can ensure that correct forms and tenses of words are being used, a Gaussian penalty based on the difference between the two sentences' length, and adding clipping to the n-gram counts, which limits the contribution of repeated n-grams once they exceed their frequency in the reference sentence. These modifications make the metric more stable while maintaining the cosine similarity framework used in CIDEr.

The Semantic Propositional Image Caption Evaluation (SPICE) metric¹³ improves the classic n-gram-based caption evaluation. This is done by explicitly modeling human consensus, while also introducing discontinuity and clipping that make the metric less vulnerable. It was also designed to address the limitations of the previously stated automatic evaluation metrics. Although SPICE was originally developed for image captioning, it can also perform video captioning tasks. In SPICE, reference and generated captions are used to create a semantic scene-graph representation that evaluates meaning rather than surface-word similarity. It focuses on whether the underlying semantic propositions in a caption correctly describe the image content.

The Semantic Textual Similarity (STS) metric¹⁴ evaluates the semantic similarity between the reference caption and the generated caption. It improves semantic textual similarity by combining Latent Semantic Analysis(LSA)-based word similarity with lexical knowledge from WordNet.¹⁵ Then, it uses this enriched word-level meaning to power alignment-based and machine learning models for scoring. This allows the system to go beyond surface-level word matching by capturing semantic relationships.

2.2 Reference-free metrics:

An alternative direction to traditional metrics is metrics based on VLMs. In this work, three different VLMs were considered, CLIP, ALIGN, and LLAMA, and are described below.

The CLIP¹⁶ model, developed by OpenAI in 2021, is defined as a dual-encoder contrastive vision-language model. It was trained on a dataset of nearly 0.4 billion image-text pairs collected from the internet.¹⁷ The CLIP

model takes an image and a caption as inputs and uses image and text encoders to retrieve visual and textual features. Then, these features are projected onto a shared embedding space, where the dot product of their embeddings creates a similarity score.¹⁸ Figure 1 illustrates the system architecture for CLIP.

The ALIGN¹⁹ model, developed by Google in 2021, is also a dual-encoder contrastive vision-language model. It was trained on a dataset of nearly 1.8 billion images and pairs of alt-text without any filtering and post-processing^{20, 21}. The ‘alt’ attribute is a parameter used in HTML web pages²² to be displayed in place of an image that cannot be displayed, which formed the basis for the training dataset. The ALIGN model uses a dual-encoder architecture (one for image and one for text). For image encoding, EfficientNet²³ is used, which is a group of image classification models. For text encoding, BERT²⁴ is used, a bidirectional transformer designed to predict tokens in a sentence. ALIGN uses these encoders to then encode images and captions into a shared vector space. ALIGN also uses contrastive learning to push away differing vectors and pull in similar vectors. This helps the model ignore irrelevant noise and information to make clearer distinctions.

The LLAMA 3.2 11B Vision model,²⁵ developed by Meta in 2025, is built on the LLAMA 3.1 text-only model, which is a large language model (LLM), and uses a separate vision adapter that feeds image encoder representations to LLAMA 3.1. The adapter was trained on text-image pairs to align image and language representations.

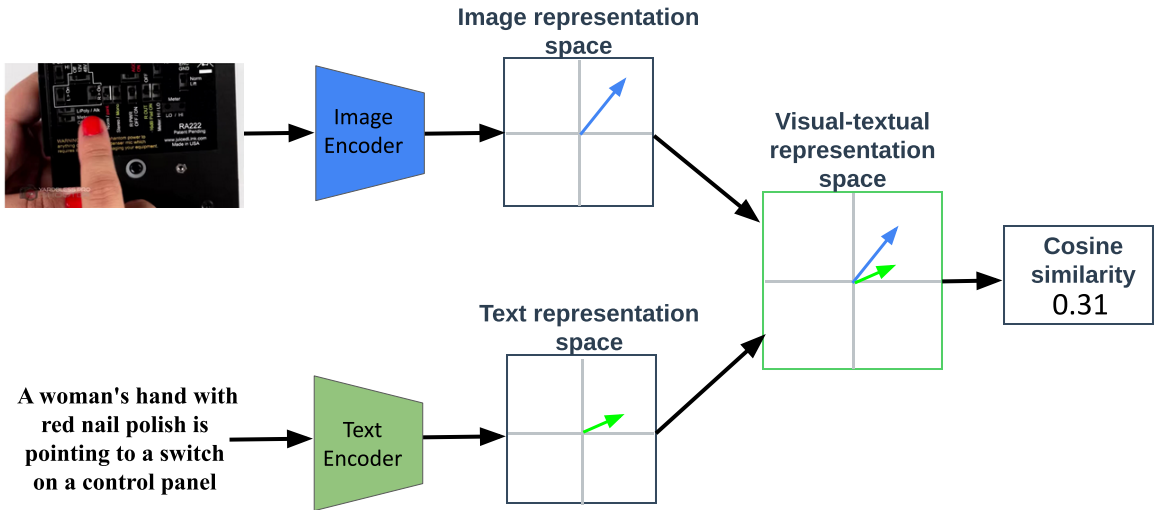


Figure 1: System architecture for the CLIP model

2.3 TRECVID benchmarking:

TRECVID⁵ (The Text Retrieval Conference - Video Retrieval Evaluation) is an annual benchmarking campaign to evaluate content-based video analysis and retrieval systems. TRECVID establishes a large-scale benchmark that is crucial to advance video analysis and captioning research. Using task-based evaluations supported by metrology, the program allows researchers to compare methods fairly. Furthermore, by testing systems annually on fixed datasets, TRECVID provides a reliable reference point for tracking long-term progress in visual tasks. As evaluation methodologies have evolved alongside these benchmark datasets, several recent surveys²⁶ also describe the rise of MLLMs and how they have changed evaluations. These models combine LLMs with multi-modal encoders, and together they can reason over multimedia. This allows them to perform caption generation, visual question answering, and multi-modal evaluation. Although all of these works are extensive, they lack exploration of reference-free evaluation on a large scale, which is addressed in this work.

3. DATA AND EXPERIMENTAL DESIGN

The data for this work was sourced from the TRECVID video-to-text (VTT) competition,²⁷ which itself is a subset of a larger dataset called V3C2 (Vimeo Creative Commons Collection 2).²⁸ This dataset included 1,757 videos and each of these videos was about 3 to 15 seconds long. To ensure diversity in the VTT dataset, 9000 videos from the V3C2 dataset were reviewed and 2000 were selected, omitting animated videos, videos that were too similar to previous videos, or videos considered inappropriate or offensive. Because of time constraints, only 1,757 videos were fully annotated and selected for this work. All ground-truth captions for the videos were produced by human annotators hired by NIST: one caption per video per annotator. In total, each video had five captions.

Teams participating in the VTT challenge were given the testing video dataset and were asked to submit one sentence describing each video focusing on four facets: Who, What, Where, and When. This is to essentially describe who in the video is doing what kind of action, and if possible, to include the location type and time or season. Each one of these videos had one caption per system submitted for the competition and one confidence score per caption, though the confidence score was not used in this analysis. There were 18 competing captioning systems, resulting in 31,626 (1,757 x 18) captions. Only a few teams utilized the audio features, as the vast majority of submitted systems relied solely on visual features.

In this work, we used three VLMs to generate scores for each of the 31,626 captions, which resulted in 94,878 (31626 x 3) individual scores. These scores were then compared to metric scores created by BLEU, METEOR, CIDER, CIDERD, SPICE, and STS. See Figure 2 for the system architecture adopted in this work.

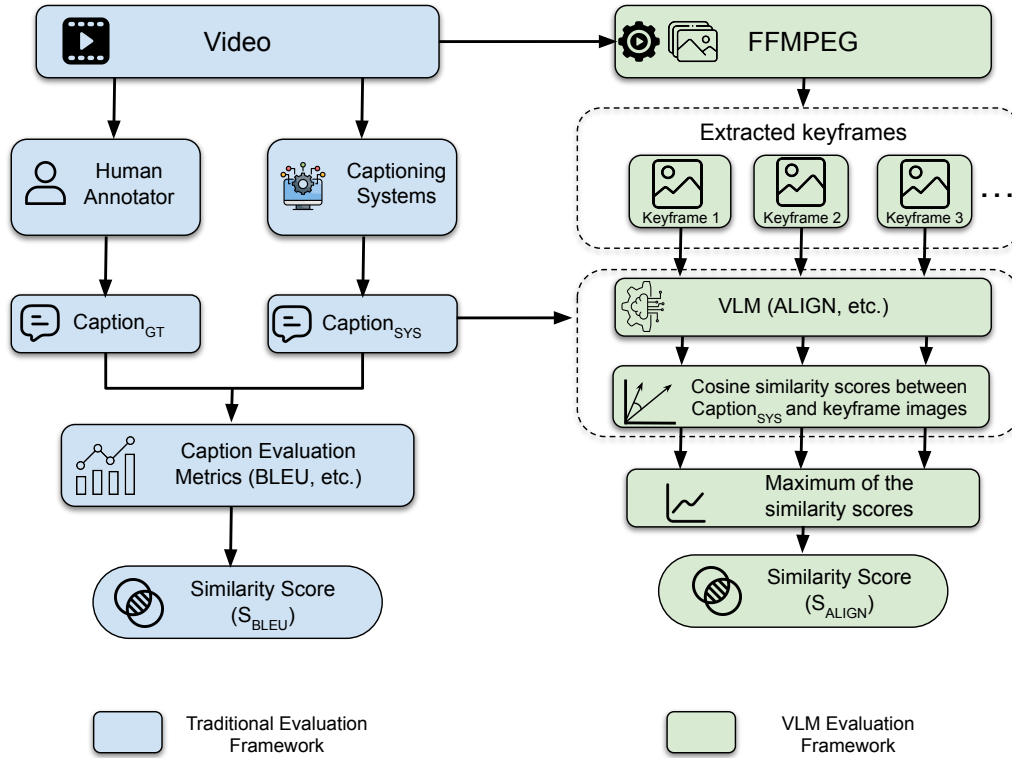


Figure 2: Evaluation framework for video captioning evaluation using traditional metrics v.s. VLM-based semantic similarity

3.1 Data preparation for use with VLMs:

To improve the throughput of this work, keyframes from the video were extracted, instead of extracting every video frame. Keyframes are complete images within a video file, whereas other frames in the video only



(a) Caption: A woman’s hand with red nail polish is pointing to a switch on a control panel (Scores: CLIP = 0.310; ALIGN = 0.131, LLAMA = 0.183)



(b) Caption: A large flock of birds are sitting on a power line (Scores: CLIP = 0.308; ALIGN = 0.209; LLAMA = 0.277)



(c) Caption: A woman paints a skull on a canvas (Scores: CLIP = 0.355; ALIGN = 0.192; LLAMA = 0.259)



(d) Caption: A man with a backpack walks down a street with a sign on his back (Scores: CLIP Score = 0.168; ALIGN = 0.156; LLAMA = 0.230)

Figure 3: Illustration of keyframes, their captions from a “Cogvlm” captioning system and similarity scores

store visual changes from these keyframes. The keyframes were extracted from the videos using FFMPEG²⁹ which resulted in about 1-6 keyframes per video, with most videos resulting in two keyframes. Additionally, the keyframe dimensions were scaled down to 224 x 224 pixels, the contrast normalized and converted to Red-Green-Blue (RGB) format. This was the requirement of the VLMs being employed in this work. These image transformations enable us to place pixel values on a standardized scale consistent with the inputs of pretrained vision model, reducing the effects of brightness and intensity differences across images.

Further, a smaller variant of LLAMA was used in this work to fit the hardware memory constraints, namely the LLAMA 3.2 11B Vision model.²⁵ This was the smallest multi-modal vision language model produced by Meta. The CLIP and ALIGN models did not require smaller variants. ALIGN was the fastest of the three VLMs in our evaluation, processed using a computer with a NVIDIA RTX 4060 GPU; CLIP was slower than ALIGN by a factor of approximately 10; LLAMA’s evaluation was approximately 100-1000 times slower than ALIGN. While LLAMA provides a general-purpose multi-modal capability, its higher inference time makes it less practical for evaluating thousands of captions. However, ALIGN’s lower inference times make it a better fit for large-scale evaluation. This suggests that when selecting a model for reference-free evaluation, one should consider both the semantic performance and computational efficiency.

3.2 Analyzing keyframes and generation of similarity scores:

The keyframe images of the video and the captions of the captioning systems were the input to the VLM, which produced similarity scores for each keyframe-caption pair (see fig. 2). To represent video-level similarity, we used the maximum similarity score across all keyframes. This helped reflect the strongest semantic match between the keyframe and the caption (the maximum is used because it captures the best semantic alignment

and avoids penalizing irrelevant or noisy frames). In the end, 1,757 similarity scores were produced for each pair of VLM-captioning systems. There were a minor fraction of cases where the VLM failed to produce any similarity scores, however, this did not influence the Pearson r values (evidenced by the low p-values corresponding to the r values).

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}} \quad (1)$$

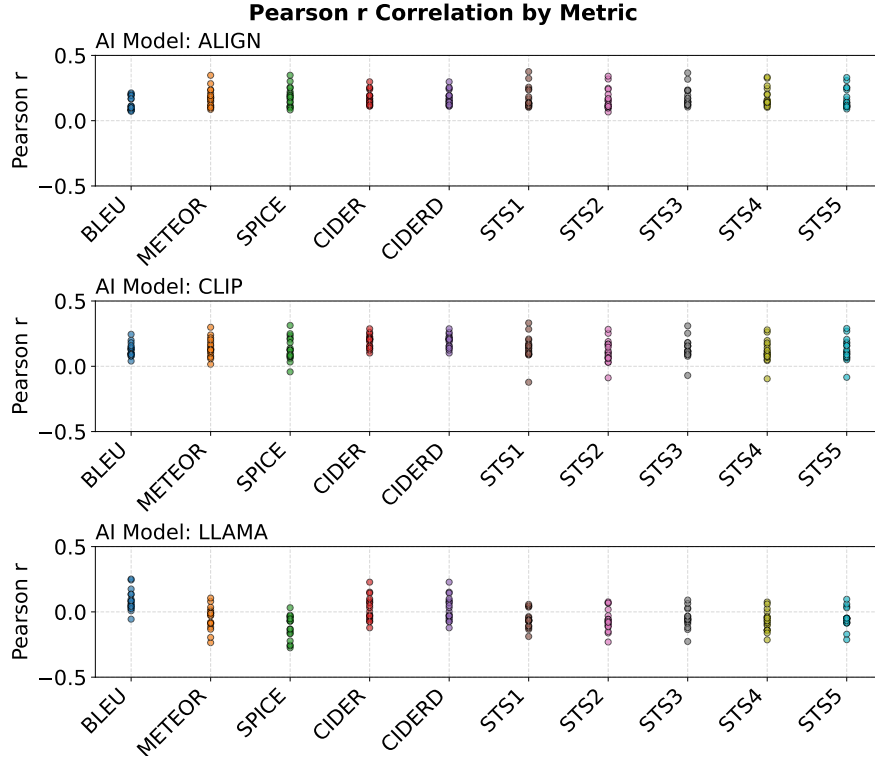


Figure 4: Pearson r correlation of traditional metrics against those of the VLM models

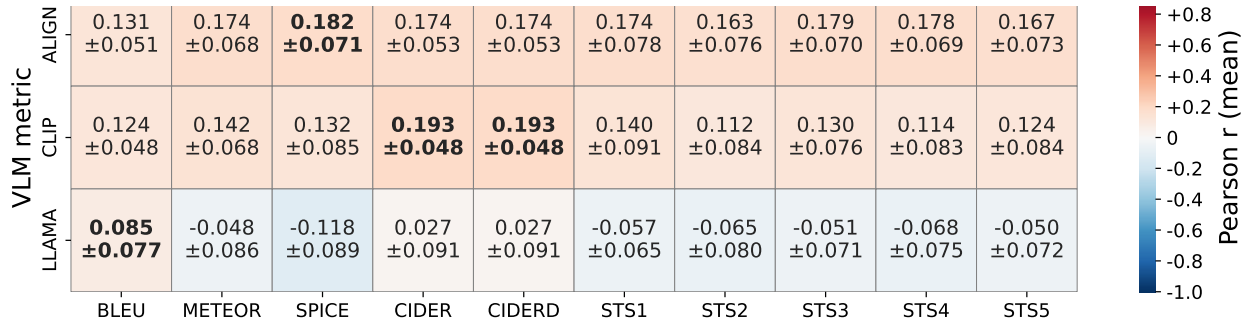


Figure 5: Heat map of mean Pearson r values of various captioning systems obtained using traditional metrics and VLMs

Table 1: Correlation between the three VLM metrics

	ALIGN	CLIP	LLAMA
ALIGN	1.000	0.778	-0.026
CLIP	0.778	1.000	0.096
LLAMA	-0.026	0.096	1.000

3.3 Comparison of similarity scores:

To ensure the reliability of the outputs from multiple VLMs, we performed manual sanity checks. These sanity checks involved manually reviewing high- and low-scoring videos and their corresponding scores. No corrective measures were performed, as all these sanity checks proved that the scores were appropriate for the videos. The correlation was calculated between system runs ($n = 18$) using the formula given in equation 1,³⁰ where x and y represent the similarity scores produced by two different methods and $\bar{x}$ and $\bar{y}$ represent the averages of the respective similarity scores.

Figure 3 shows multiple keyframes from the dataset used in this work. The caption generated by one system (Cogvlm) was used with the VLMs to generate the similarity scores, which are provided under each keyframe. Figure 4 shows the Pearson r correlation values between VLMs and the traditional metrics for captions produced by all 18 captioning systems. Figure 5 shows the same information as a heatmap. Table 1 shows the correlation between the VLM metrics for the similarity scores from all 18 captioning systems. It can be shown that ALIGN and CLIP have a strong correlation, LLAMA and CLIP show a very weak correlation, and ALIGN and LLAMA show a slight negative correlation.

4. RESULTS AND DISCUSSION

Across the three VLM-based tools, ALIGN consistently shows the strongest alignment with traditional captioning metrics, achieving moderately positive Pearson correlations ($\approx 0.072 - 0.376$) with BLEU, METEOR, SPICE, and the multiple STS metrics. The highest r values appear on the tv24_kslab submission ($r \approx 0.301 - 0.376$). One likely explanation for these findings is the underlying architecture and training objectives of the three models. ALIGN and CLIP employ similar dual-encoder architectures trained using contrastive learning to project images and text into a shared embedding space, optimizing almost identical image-text similarity objectives. Because of this, the strong correlation observed between ALIGN and CLIP in Table 1 is expected, since both models represent visual and textual information in comparable ways despite differences in scale and implementation. However, ALIGN consistently achieves stronger correlations with traditional captioning metrics than CLIP. A likely reason is the substantially larger pretraining dataset used by ALIGN. ALIGN uses a dataset of 1.8 billion image-text pairs compared to 0.4 billion image-text pairs for CLIP. This can likely produce more robust image-text similarity representations. In contrast, LLAMA shows little to no positive correlation with traditional metrics nor the other VLMS. Unlike ALIGN and CLIP, LLAMA is a general-purpose multi-modal reasoning model that uses an adapter-based architecture and is trained through a multi-stage multi-modal learning process rather than contrastive learning. Rather than learning embeddings specifically optimized for image-text similarity, LLAMA integrates visual information primarily to support language reasoning. Consequently, its weaker agreement with both traditional captioning metrics and contrastive VLMs reflects the differences in architecture and training objectives rather than a limitation of multi-modal reasoning.

These patterns hold despite missing scores in the datasets that reduce sample sizes; This indicates that all three VLM-based tools have weak correlation with traditional caption-evaluation scores, of which ALIGN is the most reliable.

Finally, the three tested VLMs varied in their correlation among themselves, with the strongest agreement between the CLIP and ALIGN models.

5. SUMMARY

In this work, we have evaluated three VLMs, namely ALIGN, CLIP, and LLAMA, using 1,757 video keyframes, captions from 18 captioning systems and 10 traditional evaluation metrics. We observed weak-to-moderate

correlations between VLM-based similarity scores and traditional evaluation metrics. Our findings suggest that while VLMs can capture aspects of semantic alignment that traditional metrics cannot capture, they are still not able to fully replace conventional approaches just yet. Moving forward, our future work will focus on refining similarity aggregation strategies and exploring vision-language model agreement frameworks to develop a more semantically grounded evaluation methodology for video captioning computational models. We will also focus on improving the robustness and reliability of VLM-based evaluation through:

- Improved frame sampling strategies.
- Incorporation of temporal modeling to better capture motion and sequence information.
- Ensemble-based evaluation using multiple VLMs to improve stability.
- Hybrid evaluation frameworks combining VLM-based semantic scoring with traditional n-gram metrics.

6. DISCLAIMER

Certain commercial equipment, instruments, software, or materials are identified in this document to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by NIST, nor necessarily the best available for the purpose. The descriptions and views contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of NIST or the U.S. Government. The National Institute of Standards and Technology Research Protections Office determined this project is “not human subjects research.”

REFERENCES

- [1] Awad, G., Curtis, K., Butt, A., Fiscus, J., Godil, A., Lee, Y., Delgado, A., Godard, E., Diduch, L., Liu, J., Graham, Y., and Quenot, G., “An overview on the evaluated video retrieval tasks at TRECVID 2022,” (June 2023).
- [2] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., Florence, P., Fu, C., Arenas, M. G., Gopalakrishnan, K., Han, K., Hausman, K., Herzog, A., Hsu, J., Ichter, B., Irpan, A., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Leal, I., Lee, L., Lee, T.-W. E., Levine, S., Lu, Y., Michalewski, H., Mordatch, I., Pertsch, K., Rao, K., Reymann, K., Ryoo, M., Salazar, G., Sanketi, P., Sermanet, P., Singh, J., Singh, A., Soricut, R., Tran, H., Vanhoucke, V., Vuong, Q., Wahid, A., Welker, S., Wohlhart, P., Wu, J., Xia, F., Xiao, T., Xu, P., Xu, S., Yu, T., and Zitkovich, B., “Rt-2: Vision-language-action models transfer web knowledge to robotic control,” (2023).
- [3] Inácio, A. d. S. and Lopes, H. S., “Evaluation metrics for video captioning: A survey,” *Mach. Learn. Appl.* **13**, 100488 (Sept. 2023).
- [4] Li, Z., Wu, X., Du, H., Liu, F., Nghiem, H., and Shi, G., “A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges,” (Jan. 2025).
- [5] National Institute of Standards and Technology, “Trecvid video retrieval evaluation,” (2026). Available at <https://trecvid.nist.gov/>. Accessed June 2, 2026.
- [6] Jurafsky, D. and Martin, J. H., [*Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*], Stanford University, Stanford, CA, 3rd edition draft ed. (2025). Accessed June 2, 2026.
- [7] Aafaq, N., Gilani, S. Z., Liu, W., and Mian, A., “Video description: A survey of methods, datasets and evaluation metrics,” *CoRR* **abs/1806.00186** (2018).
- [8] Curtis, K., Awad, G., Rajput, S., and Soboroff, I., “HLVU : A new challenge to test deep understanding of movies the way humans do,” (May 2020).
- [9] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J., “Bleu: a method for automatic evaluation of machine translation,” in [*Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*], 311–318, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002).
- [10] Lavie, A., Sagae, K., and Jayaraman, S., “The significance of recall in automatic metrics for mt evaluation,” in [*Machine Translation: From Real Users to Research*], Frederking, R. E. and Taylor, K. B., eds., 134–143, Springer Berlin Heidelberg, Berlin, Heidelberg (2004).

- [11] Banerjee, S. and Lavie, A., “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*], Goldstein, J., Lavie, A., Lin, C.-Y., and Voss, C., eds., 65–72, Association for Computational Linguistics, Ann Arbor, Michigan (June 2005).
- [12] Vedantam, R., Zitnick, C. L., and Parikh, D., “Cider: Consensus-based image description evaluation,” *CoRR* **abs/1411.5726** (2014).
- [13] Anderson, P., Fernando, B., Johnson, M., and Gould, S., “SPICE: semantic propositional image caption evaluation,” *CoRR* **abs/1607.08822** (2016).
- [14] Han, L., L. Kashyap, A., Finin, T., Mayfield, J., and Weese, J., “UMBC_EBIQUITY-CORE: Semantic textual similarity systems,” in [*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*], Diab, M., Baldwin, T., and Baroni, M., eds., 44–52, Association for Computational Linguistics, Atlanta, Georgia, USA (June 2013).
- [15] Miller, G. A., “WordNet: A lexical database for English,” in [*Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*], (1994).
- [16] Hugging Face, “Clip documentation,” (2026). Accessed June 2, 2026.
- [17] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I., “Learning transferable visual models from natural language supervision,” in [*Proceedings of the 38th International Conference on Machine Learning*], 8748–8763, PMLR (2021).
- [18] Su, C., Wang, X., Liu, S., and Chen, Y., “Efficient visual metaphor image generation based on metaphor understanding,” *Neural Process. Lett.* **56** (Apr. 2024).
- [19] Hugging Face, “Align documentation,” (2026). Accessed June 2, 2026.
- [20] Sharma, P., Ding, N., Goodman, S., and Soricut, R., “Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning,” in [*Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*], 2556–2565, Association for Computational Linguistics (2018).
- [21] Jia, C., Yang, Y., Xia, Y., Chen, Y., Parekh, Z., Pham, H., Le, Q. V., Sung, Y., Li, Z., and Duerig, T., “Scaling up visual and vision-language representation learning with noisy text supervision,” *CoRR* **abs/2102.05918** (2021).
- [22] WHATWG, “Html living standard: Requirements for providing text alternatives for images,” (2026). Accessed June 3, 2026.
- [23] Tan, M. and Le, Q. V., “Efficientnet: Rethinking model scaling for convolutional neural networks,” in [*Proceedings of the 36th International Conference on Machine Learning*], 6105–6114, PMLR (2019).
- [24] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., “Bert: Pre-training of deep bidirectional transformers for language understanding,” in [*Proceedings of NAACL-HLT 2019*], 4171–4186, Association for Computational Linguistics (2019).
- [25] Meta AI, “Llama 3.2 11b vision model card,” (2026). Accessed June 2, 2026.
- [26] Sarto, S., Cornia, M., and Cucchiara, R., “Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives,” (2025).
- [27] National Institute of Standards and Technology, “Trecvid 2020 video-to-text task guidelines,” (2020). National Institute of Standards and Technology, Gaithersburg, MD. Accessed June 2, 2026.
- [28] National Institute of Standards and Technology, “Vimeo creative commons collection (v3c2),” (2020). National Institute of Standards and Technology, Gaithersburg, MD. Accessed June 2, 2026.
- [29] FFmpeg Developers, “Ffmpeg,” (2026). Accessed June 4, 2026.
- [30] National Institute of Standards and Technology, “Engineering statistics handbook,” (2026). National Institute of Standards and Technology, Gaithersburg, MD. Accessed June 2, 2026.