Beyond Lexical Overlap: Validating Semantic Similarity Metrics for LLM-Based Video Caption Augmentation

JUVENAL FRANCISCO BARAJAS[*], San Francisco State University, USA
GIO JUNG[*], San Francisco State University, USA
MANALI SETH, San Francisco State University, USA
LANA DO, Northeastern University, USA
SAN FRANCISCO STATE UNIVERSITY, Northeastern University, USA
ANDREW SCOTT, San Francisco State University, USA
SHASTA IHORN, San Francisco State University, USA
VASSILIS ATHITSOS, University of Texas at Arlington, USA
ILMI YOON, Northeastern University, USA

Large language models (LLMs) can generate synthetic video captions at scale, offering a practical way to enrich datasets that lack multiple human references. However, evaluating the semantic reliability of such captions is challenging, particularly when traditional lexical-overlap metrics (e.g., BLEU, ROUGE, METEOR, CIDEr, SPICE) are sensitive to surface-level wording and often misalign with human judgments. This paper investigates whether sentence embedding models provide more reliable evaluation metrics for LLM-generated video captions. We first compare five embedding models—Sentence-BERT, SimCSE, MPNet, and two GPT-based embeddings—against traditional lexical metrics using 450 caption pairs from VATEX, a benchmark video captioning dataset. Through validation with a linguistic specialist, we find that sentence embedding-based approaches achieve substantially higher agreement with expert judgments using Kendall's τ correlation. We then used GPT-4o to generate ten synthetic captions per video from VATEX and apply the validated embedding metrics to compare semantic similarities with the human-human caption pairs. We applied the same pipeline to YuWa, a single-reference accessibility video dataset, which demonstrates a real-world use case. Our findings suggest that synthetic captions achieve more consistency in semantic similarity levels compared to independently written human captions. Notably, augmenting YuWa's single human caption per video with ten LLM-generated captions produces semantic similarity scores matching or surpassing those of VATEX's ten independently written human captions. This work provides the first expert-validated framework for evaluating LLM-generated video captions and demonstrates a practical pipeline for quality-controlled dataset augmentation in accessibility research.

Additional Key Words and Phrases: Sentence Embeddings, Semantic Evaluation, Data Augmentation, Video Captioning, Large Language Models

[*] These authors contributed equally to this work.

Authors' Contact Information: Juvenal Francisco Barajas, San Francisco State University, USA; Gio Jung, San Francisco State University, USA; Manali Seth, San Francisco State University, USA; Lana Do, Northeastern University, USA; San Francisco State University, Northeastern University, USA; Andrew Scott, San Francisco State University, USA; Shasta Ihorn, San Francisco State University, USA; Vassilis Athitsos, University of Texas at Arlington, USA; Ilmi Yoon, Northeastern University, USA.

Manuscript submitted to ACM

ACM Reference Format:
Juvenal Francisco Barajas, Gio Jung, Manali Seth, Lana Do, San Francisco State University, Andrew Scott, Shasta Ihorn, Vassilis Athitsos, and Ilmi Yoon. 2025. Beyond Lexical Overlap: Validating Semantic Similarity Metrics for LLM-Based Video Caption Augmentation. 1, 1 (December 2025), 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models (LLMs) have made it increasingly feasible to generate synthetic data at scale, offering a powerful solution for domains where collecting human annotations is costly, slow, or impractical. In video captioning, synthetic captions can expand existing datasets and support the training of more robust models. However, the usefulness of such synthetic data fundamentally depends on a critical question: *How do we know whether automatically generated captions are semantically reliable and sufficiently aligned with human-written descriptions?* Despite the growing use of LLM-generated captions, there is no widely accepted methodology for assessing their semantic fidelity.

Traditional automatic evaluation metrics, including BLEU [16], ROUGE [12], METEOR [4], CIDEr [21], and SPICE [3], were originally developed for machine translation or image captioning and rely heavily on lexical or syntactic overlap. These metrics perform adequately when reference and candidate captions share similar word choices, but they often fail in settings where valid descriptions differ in phrasing, focus, or level of detail. Prior analyses have shown that lexical-overlap metrics can diverge substantially from human judgments of semantic similarity [9], raising concerns about their suitability for evaluating paraphrastic or stylistically diverse captions, including those generated by LLMs.

Multi-caption video datasets, such as MSR-VTT [23], ActivityNet Captions [11], and VATEX [22], partially mitigate this issue by providing multiple human-written captions per video. In VATEX, for example, each video is associated with ten independently written captions, offering a rich representation of how different annotators describe the same visual content. However, collecting such dense annotations is expensive and often infeasible at scale. This motivates an appealing alternative: use LLMs to generate additional captions that emulate the diversity of human-written references. Even so, simply generating more captions is not enough; we must also determine whether these synthetic captions are semantically faithful to human descriptions.

Sentence embedding models offer a promising foundation for this type of evaluation. Modern embedding approaches—including Sentence-BERT [18], SimCSE [7], MPNet [19], and GPT-based embedding models [15]—encode sentences into high-dimensional vectors that capture semantic meaning beyond surface-level word overlap. These models have demonstrated strong correlation with human judgments on semantic textual similarity tasks [2, 13]. However, their use as *evaluation tools* for validating LLM-generated captions in multi-caption video datasets remains underexplored.

In this paper, we propose a principled pipeline for evaluating LLM-based caption augmentation using sentence embedding models, grounded in expert human judgment. We focus primarily on the VATEX dataset as a controlled testbed. First, we construct a baseline from ten VATEX videos, each with ten human-written captions, yielding 450 caption pairs. We compute similarity scores for these pairs using five sentence embedding models and several traditional lexical metrics, and compare each metric's ranking against judgments from a linguistic expert using Kendall's τ correlation [8]. This analysis identifies which similarity measures best align with human semantic expectations. We then use the validated embedding-based metrics to evaluate LLM-generated captions for the same VATEX videos, where one human caption is paired with ten synthetic captions per video, allowing us to compare the semantic similarity distributions of human–human and human–LLM caption sets.

Finally, we illustrate the broader applicability of this pipeline on YuWa [17], a real-world video description dataset that provides only a single human caption per video. By generating additional captions with an LLM and evaluating them with the validated embedding metrics, we demonstrate how the same methodology can be used to assess the quality of synthetic augmentation in low-resource, single-reference settings.

Our contributions are threefold:

- We provide a systematic comparison of sentence embedding models and classical lexical metrics as tools for measuring semantic similarity between video descriptions, validated against expert linguistic judgments on VATEX.
- We demonstrate that embedding-based similarity metrics align more closely with human semantic judgments than traditional overlap-based metrics, making them more suitable for evaluating LLM-based caption augmentation.
- We present a replicable pipeline for using LLMs to augment caption datasets and for validating the semantic reliability of the resulting synthetic captions, with VATEX as a controlled benchmark and YuWa as a real-world use case.

Taken together, these results offer practical guidance for researchers who wish to augment video captioning datasets with LLM-generated descriptions while maintaining control over semantic quality.

2 Related Work

Research on evaluating similarity for video descriptions spans classical lexical-overlap metrics, sentence embedding models, large-scale captioning datasets, and recent developments in LLM-driven data augmentation. This section reviews these areas and highlights how they have been used to support video–language research.

2.1 Lexical and N-gram Based Evaluation Metrics

A large body of captioning research has relied on lexical- or *n*-gram–based evaluation metrics. BLEU [16] measures precision over *n*-grams and was originally introduced for machine translation. ROUGE [12] was developed for summarization and emphasizes recall over *n*-gram matches. METEOR [4] partially mitigates strict lexical mismatch by incorporating stemming and synonym matching. For image and video captioning, CIDEr [21] emphasizes consensus among multiple human descriptions by weighting *n*-grams with TF-IDF, and SPICE [3] evaluates semantic relations via scene-graph tuples.

Despite their widespread use, these metrics often correlate imperfectly with human judgments, especially when captions differ in style, specificity, or phrasing. Kilickaya et al. [9] show that widely adopted captioning metrics behave inconsistently across datasets and linguistic variations, suggesting that *n*-gram overlap is insufficient for capturing semantic similarity in many real-world scenarios. As video–language models and generation systems increasingly produce diverse paraphrases, interest has grown in evaluation methods that better reflect sentence-level meaning beyond token overlap.

2.2 Sentence Embeddings and Semantic Similarity

Sentence embedding models provide vector representations that capture semantic content and have become a common tool for measuring similarity between sentences. Early work such as Skip-Thought [10] explored sequence-prediction–based encoders, while contextualized models such as BERT [5] greatly improved representation quality.

Sentence-BERT [18] introduces a Siamese architecture optimized for semantic textual similarity (STS), enabling more accurate similarity scoring. SimCSE [7] uses contrastive learning to further strengthen semantic discrimination, and MPNet [19] combines masked and permuted language modeling to improve contextual encoding. GPT-based embedding models [15] extend this trend with strong performance on text similarity and retrieval tasks.

Large-scale STS evaluations such as SemEval [2, 13] show that embedding-based methods correlate strongly with human similarity ratings. These models have therefore been adopted in various captioning, retrieval, and multimodal alignment tasks where semantic proximity is important. Their use as evaluation metrics for generated captions is increasingly explored as researchers seek alternatives to classical lexical metrics in contexts involving paraphrasing or stylistic variation.

2.3 Video Captioning Datasets

Benchmark video captioning datasets differ significantly in the diversity and quantity of human-provided descriptions. MSR-VTT [23] pairs short video clips with multiple crowd-sourced captions, while ActivityNet Captions [11] provides temporally localized descriptions for long videos. VATEX [22] offers ten independently written captions per video, making it useful for studying linguistic variation, training captioning models, and evaluating multi-reference metrics.

Outside benchmark settings, many video datasets contain only a single human-written caption per clip. YuWa [17], for example, compiles community-authored descriptions originally designed for video accessibility platforms. Although the dataset originates in an accessibility context, it is representative of a broader class of single-reference, low-resource video captioning corpora where the lack of multiple human descriptions limits both training diversity and evaluation stability. Such datasets highlight the practical need for scalable methods to enrich or diversify caption sets.

2.4 LLM-based Data Augmentation

Large language models have recently been explored as tools for generating synthetic training data across a range of NLP and multimodal tasks. Surveys such as Ding et al. [6] document approaches for paraphrasing, elaboration, and controlled text generation for data augmentation. In the video domain, LLMs have been used to enhance temporal grounding datasets and reduce annotation requirements by producing diverse language queries and alternative expressions [20].

Video captioning datasets that contain only a single reference caption—such as YuWa and other domain- or platform-specific collections—present a natural application area for LLM-based augmentation. Generating multiple captions per clip can approximate the multi-reference structure of datasets like VATEX, where linguistic diversity supports more robust model training and evaluation. However, LLM-generated captions often vary in style, detail, or lexical choice relative to the original human caption, making evaluation with traditional n-gram metrics less reliable. This has motivated growing interest in semantic similarity measures, including sentence embeddings, as tools for assessing the quality and consistency of synthetic captions produced through LLM-driven augmentation workflows.

3 Assessing the Reliability of Sentence Embedding Models

3.1 Baseline Dataset analysis (VATEX)

For this initial experiment, ten videos were randomly selected from the VATEX dataset. Each selected video is associated with 10 human-written captions, resulting in $\binom{10}{2} = 45$ unique sentence pairs per video when all possible pairwise combinations are considered. In total, this yielded 450 sentence pairs for evaluation. While this represents a relatively small subset of the full dataset, it is already substantial, particularly in preparation for subsequent expert-based annotation, where each pair requires careful human judgment. Thus, selecting ten videos represents a balanced trade-off between experimental feasibility and analytical value.

These 450 sentence pairs form the baseline against which different sentence embedding models are evaluated. Since all captions within a video are written to describe the same visual sequence, they are expected to exhibit a meaningful degree of semantic relatedness, though not strict paraphrasing. This characteristic makes the dataset particularly suitable for testing the sensitivity and reliability of embedding-based similarity measures.

Table 1. Mean and standard deviation of similarity and lexical overlap metrics on the VATEX baseline with 450 sentence pairs.

	Sentence-BERT		**SimCSE**		**MPNet**		**gpt-emb3-small**		**gpt-emb3-large**		**CIDEr**		**CIDErD**	
Dataset	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
VATEX	0.6166	0.1588	0.5847	0.1405	0.5374	0.1587	0.5446	0.1324	0.5426	0.1329	0.5715	0.8525	0.4983	0.8048
	BLEU-1		**BLEU-2**		**BLEU-3**		**BLEU-4**		**SPICE**		**METEOR**		**ROUGE**	
Dataset	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
VATEX	0.2818	0.1234	0.1274	0.1272	0.0380	0.0980	0.0143	0.0717	0.1541	0.1396	0.1422	0.0714	0.2670	0.1119

As shown in Table 1, a clear separation emerges between the two groups of evaluation methods. The sentence embedding models—Sentence-BERT, SimCSE, MPNet, and the two OpenAI embedding models—consistently produce higher mean similarity scores with relatively constrained variance. In contrast, the traditional n-gram and lexical-overlap-based metrics, including CIDEr, BLEU, METEOR, ROUGE, and SPICE, exhibit substantially lower mean scores and, in some cases, significantly larger variability. This pattern suggests that sentence embedding approaches are better suited for capturing semantic relatedness between different descriptions of the same visual content, while traditional metrics are highly sensitive to surface-level lexical differences. However, a higher numerical similarity score does not inherently imply that a metric is reliable or semantically aligned with human judgment. To determine whether these embedding-based similarity measures meaningfully reflect human perceptions of sentence relatedness, an expert-driven validation is required. Therefore, the next step in our analysis compares the similarity rankings produced by these models with those provided by a linguistic expert.

3.2 Alignment with Linguistics Expert

To contextualize and validate the embedding-based similarity scores obtained from the VATEX baseline subset, we consulted a linguistic expert who holds expertise in computational linguistics and the syntax-semantics interface, with a research background in large language model evaluation and extensive experience in multimodal language education and curriculum design for diverse learner populations. The expert independently assessed the semantic relatedness of the same 450 caption pairs using the Described and Captioned Media Program (DCMP) guidelines [1] which provides well-defined and widely cited guidance for educational and instructional contexts, emphasizing precision and consistency in learning environments. From this, the expert derived three discrete, ordinal similarity scales bounded between 0 and 1: a binary scale, a 5-level scale in 0.25 increments, and a 9-level scale in 0.125 increments. The highest level corresponded to clear semantic equivalence (i.e., the sentences expressed the same meaning), the intermediate level captured partial semantic overlap (i.e., the sentences shared related meaning but differed in focus or specificity), and the lowest level indicated minimal or weak semantic similarity. This structured yet flexible scale was chosen to reflect the nuanced judgments involved in human interpretation of sentence meaning.

To quantify the degree of agreement between the expert's judgments and the automated similarity scores, Kendall's τ rank correlation coefficient [8] was computed for each metric. The figure 1 presents the Kendall's τ values for each

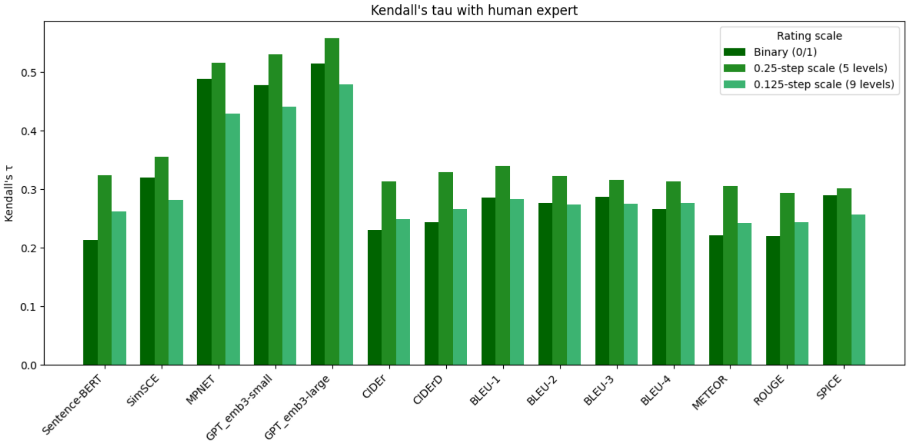

Fig. 1. Kendall's τ correlation result with the human expert across sentence embedding models and traditional lexical-overlap metrics.

metric in comparison with expert judgments. Across all comparisons, the sentence embedding models consistently exhibited higher correlation with the linguistic expert's ratings than the traditional NLP metrics. This suggests that embedding-based approaches are better aligned with human perception of sentence-level semantic similarity, as they capture meaning beyond surface-level word overlap.

Overall, the strong alignment between expert judgments and sentence embedding scores provides empirical support for the use of embedding-based similarity as a reliable measurement tool. This validation is a critical step before applying the same methodology to evaluate augmented datasets and alternative data sources in subsequent experiments.

4 Data Augmentation on VATEX

Having established that sentence embedding models align more closely with expert judgments than traditional lexical metrics on human–human VATEX captions, we next examine how these models behave when applied to LLM-generated captions. The goal of this experiment is to assess whether synthetic captions produced by an LLM can achieve semantic similarity levels comparable to those observed among independently written human captions, using the same five embedding models validated in the previous section.

4.1 Prompting GPT

For the same 10 VATEX videos that were used in the baseline analysis, a single human-written caption was randomly chosen to serve as a semantic anchor. This original caption, together with the corresponding eight visual frames, was provided to OpenAI's GPT 4o API [14] using the system prompt (as shown in Figure 2) to generate 10 distinct captions. Importantly, this prompt was designed using the same DCMP guidance that informed the linguistic expert's evaluation criteria, ensuring consistency in the conceptual standard of a "high-quality" description across both human and model-based judgments.

```
system_prompt = f"""Generate 10 unique, distinct, and concise descriptions for a video segment and text
from a human describer. These descriptions should be tailored to be accessible for blind and visually
impaired individuals following these guidelines.

Guidelines - A Quality Description Must Be:

Accurate: There must be no errors in word selection, pronunciation, diction, or enunciation. Prioritized:
Content essential to the intended learning and enjoyment outcomes is of primary importance.
Consistent: Both the description content and the voicing should match the style, tone, and pace of the
program.
Appropriate: Consider the intended audience, be objective, and seek simplicity and succinctness.
Equal: Equal access requires that the meaning and intention of the program be conveyed.

The 8 frames are provided as a single image where frames are read from left to right. The original text
description follows. Please make each description representative of the whole video clip and not frame
by frame. Do not give titles to each descrpiton. Seperate each description with two newline characters.
Do not print the number of the description.
"""
```

Fig. 2. System prompt used for generating video captions.

4.2 Sentence Embedding on augmented VATEX dataset

We next applied the five validated sentence embedding models to the augmented VATEX captions generated in the previous subsection. For each of the ten videos, all pairwise combinations among the eleven captions (1 human + 10 LLM-generated) were computed, and cosine similarity was used to quantify semantic similarity for each sentence pair. This resulted in a total of 550 sentence pairs, evaluated using the same procedure as in the baseline human–human analysis.

Table 2. Mean and standard deviation of similarity scores for embedding-based metrics on the augmented VATEX with 550 sentence pairs.

	Sentence-BERT		**SimCSE**		**MPNet**		**gpt-emb3-small**		**gpt-emb3-large**	
Dataset	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
VATEX	0.7932	0.0995	0.7252	0.1149	0.7500	0.0981	0.7245	0.0852	0.7451	0.0810

Table 2 reports the mean and standard deviation of similarity scores for the augmented VATEX dataset. Across all five models, the mean similarity scores are noticeably higher than those observed in the original VATEX baseline (Table 1), where all captions were human-written. At the same time, the standard deviations are generally lower, indicating more concentrated similarity distributions within each video.

This pattern suggests that the LLM-generated captions are not only semantically comparable to the original human description, but also more closely aligned with one another than independently written human captions. While human annotators naturally introduce diverse perspectives, levels of detail, and linguistic style, the LLM is guided by a single semantic anchor and a consistent prompt structure, which may lead to more uniform representations of the same content. Importantly, the high mean similarity values alone do not establish that these captions are “better” or “more correct,” but they do indicate that the augmented captions form a semantically coherent cluster around the human reference description.

These results provide initial quantitative evidence that LLM-generated captions, guided by a structured prompt, can achieve a more consistent level of semantic similarity compared to human-written captions in benchmark datasets such as VATEX. This finding motivates the application of the same augmentation and evaluation framework to real-world, single-caption datasets such as YuWa in the following section.

5 Application on YuWa Datset

The previous sections established that (i) sentence embedding models align more closely with expert judgments than traditional lexical metrics on human–human VATEX captions, and (ii) LLM-generated captions on VATEX form semantically coherent clusters around a human anchor caption. We now examine whether this evaluation and augmentation pipeline transfers to a real-world use case with YuWa (need to modify here).

To construct a comparable evaluation setting, we randomly sampled 100 videos from YuWa and applied the same augmentation procedure used for VATEX. For each video, we selected the original human description as the semantic anchor and used the DCMP-informed prompt (Figure 2) to generate ten additional captions with OpenAI's GPT 4o API [14]. This yielded eleven captions per video, and all pairwise combinations among them (55 pairs per video) were scored using cosine similarity under the same five sentence embedding models as before.

For a fair comparison, we also constructed a 100-video augmented subset of VATEX using the same protocol: one human caption plus ten LLM-generated captions per video, and 55 pairwise similarities per video. Table 3 summarizes the mean and standard deviation of similarity scores for both datasets.

Table 3. Mean and standard deviation of similarity scores for embedding-based metrics on augmented VATEX (100 randomly selected videos) and augmented YuWa (100 randomly selected videos). Each dataset yields 5,500 sentence pairs (100 videos × 55 pairs).

	Sentence-BERT		**SimCSE**		**MPNet**		**gpt-emb3-small**		**gpt-emb3-large**	
Dataset	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
VATEX	0.8536	0.0784	0.7548	0.0943	0.7929	0.0810	0.7587	0.0715	0.7694	0.0689
YuWa	0.8350	0.1769	0.7287	0.2006	0.7739	0.1843	0.7688	0.1722	0.7777	0.1763

The mean similarity scores for YuWa are closely aligned with those of augmented VATEX across all five embedding models, with differences on the order of a few hundredths. This suggests that, on average, the LLM-generated captions for YuWa are semantically close to their human anchors as those generated for VATEX, despite differences in domain, narration style, and source platform.

A key difference appears in the standard deviations: YuWa exhibits substantially higher variance than VATEX for all models. This is a plausible consequence of greater heterogeneity in YuWa content and narration practices. While VATEX captions are collected under controlled annotation protocols, YuWa descriptions are created by a diverse set of volunteers, with varying levels of detail, style, and adherence to audio description guidelines. When LLMs are anchored on this more heterogeneous human input, the resulting synthetic caption sets naturally span a wider range of similarity values.

These results support two claims. First, the sentence embedding models that were validated against expert judgments on VATEX behave consistently when applied to LLM-augmented captions in a real-world setting. Second, the mean similarity levels observed for YuWa closely track those of a multi-caption benchmark, providing evidence that LLM-based augmentation can approximate the semantic density of datasets like VATEX even when starting from a single human description per video. This positions our pipeline as a practical tool for researchers who wish to expand video captioning datasets while maintaining a quantitative handle on the semantic reliability of the augmented captions.

6 Limitations and Future Work

This study establishes a methodology for evaluating LLM-generated video captions using sentence embeddings validated against expert judgment. Several limitations suggest directions for future research.

First, the scale of our dataset and expert validation is limited, with only one expert evaluating 450 sentence pairs from ten VATEX videos. While this sample size provides meaningful insight into metric alignment with human semantic judgments, expanding the validation to include more videos, additional experts, and inter-rater reliability analysis would strengthen then generalizability of our findings. Future work should incorporate multiple linguistic experts with diverse backgrounds to capture a broader range of semantic interpretations and reduce potential individual bias.

Second, our augmentation pipeline relies on a single prompt design informed by the DCMP guidelines. Alternative prompting strategies, including few-shot examples, chain-of-thought reasoning, or iterative refinement approaches, may yield captions with different semantic characteristics. Systematic exploration of prompt engineering techniques and their impact on caption quality represents a valuable direction for future investigation.

Third, generating multiple captions per video using LLMs incur computational and financial costs that may limit scalability for very large datasets. Future research should investigate cost-performance trade-offs by systematically comparing propriety and open-source LLMs, analyzing the marginal semantic benefit of generating additional captions per video, and evaluating selective augmentation approaches. Understanding these trade-offs would enable researchers to make informed decisions about augmentation scale based on their specific budget and quality requirements.

Finally, our current pipeline generates captions based solely on visual frames and a single human reference description, without incorporating additional modalities such as audio tracks, temporal dynamics, or contextual metadata. Future work should explore multimodal integration approaches that leverage these additional information sources to enhance the semantic accuracy and richness of LLM-generated captions. Systematic investigation of multimodal prompt designs represents a promising avenue for improving augmentation quality and better capturing the full complexity of video content.

Addressing these limitations systematically represents our roadmap for advancing this work. This study establishes critical foundational contributions: the first expert-validated framework for evaluating LLM-generated video captions, empirical evidence of their semantic reliability, and a replicable pipeline that addresses the longstanding challenge of scalable dataset augmentation in accessibility-focused video description research.

7 Conclusion

This study presents a scalable framework for augmenting video captioning datasets using LLMs and validating the semantic reliability of the resulting captions through sentence embedding models. Using VATEX as a benchmark, we show that embedding-based similarity measures align more closely with expert judgments, grounded in DCMP guidelines, than traditional lexical metrics. This establishes sentence embeddings as a principled tool for assessing semantic fidelity in caption generation tasks.

When applied to LLM-generated captions, the embedding models produced higher and more stable similarity scores than those observed among independently written human captions in VATEX. These findings indicate that, when guided by a structured prompt, LLMs can generate semantically coherent caption sets that remain closely aligned with a human reference. Rather than replacing human descriptions, this process enables controlled, semantically consistent expansion of existing datasets.

Applying the same pipeline to YuWa dataset demonstrates its viability in real-world, low-resource settings. Despite increased variance in community-authored descriptions, the overall similarity patterns remained consistent with those observed in augmented VATEX. Our results suggest a practical strategy for dataset expansion, where LLMs reduce annotation cost and effort and sentence embedding models act as a reliability filter to maintain semantic quality in video understanding applications.

References

[1] 2024. Described and Captioned Media Program (DCMP). https://dcmp.org/learn/descriptionkey

[2] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*. 497–511.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*. Springer, 382–398.

[4] 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*. 65–72.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[6] Bosheng Ding, Chengwei Qin, Ruochen Zhao, Tianze Luo, Xinze Li, Guizhen Chen, Wenhan Xia, Junjie Hu, Anh Tuan Luu, and Shafiq Joty. 2024. Data augmentation using large language models: Data perspectives, learning paradigms and challenges. *arXiv preprint arXiv:2403.02990* (2024).

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).

[8] Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika* 30, 1-2 (1938), 81–93.

[9] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 199–209.

[10] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems* 28 (2015).

[11] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*. 706–715.

[12] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.

[13] Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 1–8.

[14] OpenAI. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276* (2024). https://arxiv.org/abs/2410.21276 Version v1, 25 Oct 2024.

[15] OpenAI. 2024. New OpenAI Embedding Models. https://openai.com. Accessed: 2025-11-23.

[16] 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*. 311–318.

[17] Charity Pitcher-Cooper, Manali Seth, Benjamin Kao, James M Coughlan, and Ilmi Yoon. 2024. You Described, We Archived: A rich audio description dataset. In *Journal on technology and persons with disabilities:... Annual International Technology and Persons with Disabilities Conference*, Vol. 11. 192.

[18] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[19] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems* 33 (2020), 16857–16867.

[20] Yun Tian, Xiaobo Guo, Jinsong Wang, and Bin Li. 2025. Enhancing video temporal grounding with large language model-based data augmentation. *The Journal of Supercomputing* 81, 5 (2025), 658.

[21] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4566–4575.

[22] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4581–4591.

[23] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.