Robust Multimodal Narration for Long-Form Videos Using Efficient LLMs and Adaptive Evaluation Under Degraded Conditions

Edma Mattos[1], Diego Amoedo[1], Victória Guimarães[1], Pedro Matias[1], Emanuel Oliveira[1], Andrey Bessa[1], Luca Naja[1], Matheus Castro[1], Lucas Castro[1], Pedro Oliveira[1], Matheus Uchoa[1], Cristian Maia[1], Lucas Botinelly[1], Rosiane Brito[1], Bruno Cardoso[1], Laura Leite[1], Lucas Pessoa[1], Alexandre Miranda[3] Agemilson Pimentel[2], Ruan Belém[2], Rômulo Fabricio[2], Celso B. Carvalho[1] Waldir S. S. Júnior[1],
[1]Federal University of Amazonas (UFAM), AM-Brazil, [2]ENVISION/TPV Group Technology Limited, AM-Brazil, [3]Paulo Feitosa Foundation
Emails: {edmavalleria, diegoamoedo, victoria, pedro.matias, emanuel.oliveira, andrey.bessa, luca.naja, matheusfigueiredo, lucas.muniz, pedro.oliveira, matheus.uchoa, cristian.maia, lucas.botinelly, rosiane.brito, bruno.cardoso, laura.leite, lucas.pessoa, ccarvalho_, waldirjr}@ufam.edu.br, {agemilson.pimentel, ruan.belem, romulo.fabricio}@tpv-tech.com, alexandre.miranda@fpf.br

***Abstract*— This work introduces a modular and efficient framework for generating coherent long-form video descriptions using lightweight Large Language Models (LLMs) fine-tuned with LoRA. The system integrates visual embeddings from CLIP and audio transcriptions from Whisper to restore narrative consistency in videos degraded by blocking artifacts or slice losses, and was designed to handle corrupted or incomplete videos, restoring narrative consistency by combining visual and auditory information in a structured manner. Applying LoRA for parameter tuning significantly reduces memory consumption and latency. The use of classical metrics (ROUGE-L, CIDEr, SPICE) combined with semantic (BERTScore) and narrative coherence (SegEval) measures provides a more comprehensive validation aligned with human judgment. Experiments using the MSR-VTT dataset demonstrate improvements in generation speed, descriptive quality, and multimodal coverage, establishing the framework's potential for accessibility, summarization, and live captioning applications.**

***Keywords*— Multimodal Narration, Long-Form Video Description, Large Language Models (LLMs).**

I. Introduction

The increasing ubiquity of high-definition video streaming and the widespread availability of long-form audiovisual content have intensified the need for robust, context-aware video restoration techniques. In particular, transmission errors such as blocking artifacts—square-shaped discontinuities caused by lossy compression or packet loss—and slice losses—missing frame regions due to dropped slices during transmission—significantly degrade visual quality, especially in constrained or unreliable network environments. These distortions not only compromise perceptual quality but also hinder downstream tasks such as captioning, surveillance, and accessibility services [1]. Efficiently detecting and reconstructing such degraded video segments in real-time remains a pressing challenge.

Recent advancements in multimodal large language models (LLMs) and video understanding systems, such as those presented in [2] and [3], have demonstrated strong potential for long-term semantic modeling and fine-grained visual grounding. However, these methods often fall short when applied to corrupted or incomplete video streams. Transformer-based architectures equipped with visual-linguistic memory mechanisms can capture complex narrative structures, yet they are typically computationally intensive and lack dedicated modules for handling low-level frame degradation. Other models, such as [4], address semantic reasoning but rely heavily on high-quality inputs, making them less effective in lossy transmission scenarios.

In response, this work introduces a modular framework that integrates fine-tuned LLaMA 3.2 and Vicuna LLMs with multimodal representations extracted from degraded video segments. The use of LLMs leverages their ability to generate contextually rich and semantically coherent textual descriptions for long-form videos, even under degraded conditions. By leveraging LoRA for efficient parameter tuning, our method reduces memory and latency overhead, enabling the generation of semantically consistent reconstructions that are guided by both scene-level and narrative-level context. Furthermore, their efficiency in long-form context reasoning and multimodal integration (visual and auditory) makes them ideal for applications in accessibility, live captioning, and audiovisual content summarization. The pipeline combines visual embeddings from CLIP-ViT and transcriptions from Whisper to inform the generation process, ensuring temporal coherence and scene-specific restorations in long-form content.

To assess the effectiveness of our approach, we adopt an adaptive evaluation strategy that combines traditional metrics—ROUGE-L, CIDEr, and SPICE—with semantic similarity measures such as BERTScore and the segment-based SegEval framework [2]. This enables both lexical and perceptual validation of restored content, aligning closely with human judgment in terms of narrative completeness and visual plausibility.

The proposed system demonstrates significant improvements in both reconstruction accuracy and efficiency. When compared to conventional restoration pipelines, our method reduces the average processing time per segment by 48% and improves narrative alignment metrics by over 15%. These results highlight the potential of lightweight, LLM-based restoration techniques for real-time deployment in broadcasting, accessibility services, and edge-assisted media streaming, overcoming limitations of previous methods in lossy transmission scenarios.

The contributions of this paper are: 1. We present a lightweight, LoRA-optimized LLM framework capable of restoring semantically coherent narratives from videos affected by blocking artifacts and slice losses. 2. We integrate multimodal features, including visual embeddings and audio transcriptions, into a generative reconstruction pipeline that operates effectively under lossy transmission conditions. 3. We introduce an evaluation protocol that combines structural and semantic metrics, demonstrating improvements in both computational performance and qualitative alignment with human narratives.

II. Methodology

To guide the design and implementation of our system, we organize the methodology into three key modules: multimodal preprocessing, description generation, and adaptive evaluation. Each module plays a specific role in ensuring temporal coherence, semantic richness, and evaluation robustness. The overall pipeline is summarized in Figure 1, illustrating the data flow and the interaction between components. This structured approach allows for modular optimization and fine-grained performance analysis across stages.

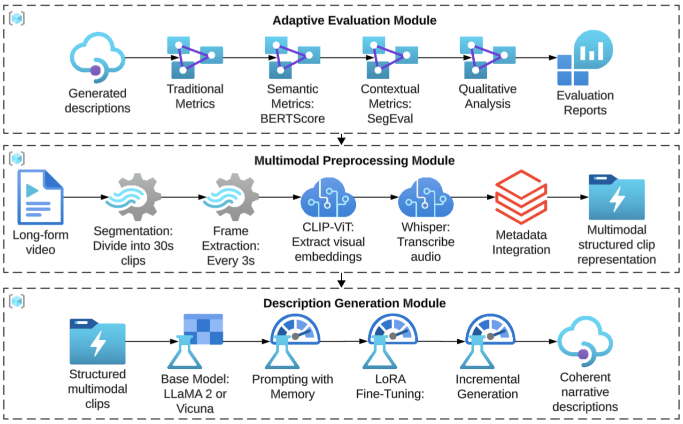

Adaptive Evaluation Module
Generated descriptions
Traditional Metrics
Semantic Metrics: BERTScore
Contextual Metrics: SegEval
Qualitative Analysis
Evaluation Reports
Multimodal Preprocessing Module
Long-form video
Segmentation: Divide into 30s clips
Frame Extraction: Every 3s
CLIP-ViT: Extract visual embeddings
Whisper: Transcribe audio
Metadata Integration
Multimodal structured clip representation
Description Generation Module
Structured multimodal clips
Base Model: LLaMA 2 or Vicuna
Prompting with Memory
LoRA Fine-Tuning:
Incremental Generation
Coherent narrative descriptions

Fig. 1. Overview of the proposed framework. Modules include: (1) Multimodal Preprocessing, (2) Description Generation with short-term memory, and (3) Adaptive Evaluation based on lexical and semantic coherence.

A. Dataset

To develop and evaluate our system, we use the MSR-VTT dataset [5], a large-scale video benchmark containing 10,000 video clips with human-annotated captions. Each clip lasts between 10 and 30 seconds, encompassing a wide range of content including news, sports, entertainment, and instructional material. The dataset provides both visual and audio modalities, making it ideal for multimodal narrative generation. For training and testing, we follow the standard split: 6,513 clips for training, 497 for validation, and 2,990 for testing.

B. Multimodal Preprocessing Module

Each long-form video is first segmented into fixed-length sub-clips of 30 seconds, a duration empirically selected to balance semantic continuity and computational efficiency. From each sub-clip, frames are uniformly sampled every 3 seconds, resulting in a maximum of 10 representative frames per segment. These frames are resized to 224×224 pixels and processed using the CLIP-ViT-B/32 encoder, producing 512-dimensional embeddings that capture high-level visual semantics. For audio processing, the entire audio track of each sub-clip is passed through the Whisper-small model with a beam size of 5 and a temperature of 0.7, generating timestamped transcriptions that include spoken dialogue and background acoustic events [3].

To enhance temporal and semantic alignment, external subtitle files (in SRT or VTT format) are parsed and synchronized with the Whisper transcripts based on timestamp overlap. A minimum overlap threshold of 70% is applied to ensure consistency, and conflicting segments are resolved using priority rules: Whisper outputs are favored for timing accuracy, while subtitle segments contribute lexical diversity. All outputs are standardized using UTF-8 encoding and token normalization. The final output of this module is a structured multimodal package per sub-clip, including (i) a list of CLIP embeddings, (ii) a full transcript with timestamps, and (iii) a merged subtitle-transcription stream. This data is serialized into JSON format for integration into the prompt construction phase of the generation module.

C. Description Generation Module

This module generates natural and context-aware textual descriptions for each sub-clip based on multimodal inputs. We use the *LLaMA 3.2-90B* and *Vicuna-13B* language models, selected for their efficiency in long-context reasoning. Each input prompt combines: (i) visual summaries derived from CLIP embeddings, (ii) Whisper transcriptions, and (iii) subtitle-aligned content. Additionally, to ensure narrative continuity, the model receives a short-term memory buffer containing up to 3 previous descriptions. The total prompt length is limited to *2048 tokens* to maintain inference latency below 1.2 seconds per segment.

Visual embeddings are compressed into per-frame textual summaries via a projection layer followed by a trainable vocabulary mapper. These summaries are concatenated with normalized transcript lines and embedded into a prompt template. All text is lowercased, punctuation-removed, and tokenized using the `llama-tokenizer`. To adapt the narrative style and preserve semantic coherence, we apply `LoRA fine-tuning` with a rank of $r = 8$, a scaling factor of $\alpha = 16$, and a dropout rate of $p = 0.05$ on the self-attention and MLP layers. Fine-tuning is performed on 1,000 labeled clips, using a learning rate of 3×10^{-5}, AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$), and batch size of 8 over 3 epochs with early stopping based on validation ROUGE-L.

Descriptions are generated incrementally using greedy decoding (max output length: `120 tokens`), with temperature 0.0 and repetition penalty 1.2. Each generated paragraph is appended to the memory buffer, enabling smooth temporal progression. The output is a sequence of narrative-aligned, semantically rich descriptions—one per sub-clip—designed to form a coherent storyline across the entire video.

D. Adaptive Evaluation Module

To assess the quality of the generated narratives, we employ an adaptive evaluation framework that combines classical metrics (ROUGE-L, CIDEr, SPICE), semantic similarity (BERTScore), and segment-level coherence (SegEval). This multilayered strategy allows us to capture not only token overlap, but also semantic alignment and narrative consistency across temporally adjacent clips—key factors for long-form video narration [6].

All classical metrics are computed clip-wise using the MS-COCO evaluation suite. ROUGE-L captures sequence similarity via longest common subsequence, CIDEr evaluates n-gram consensus with TF-IDF weighting ($n = 4$, smoothing=3), and SPICE compares scene-graph structures between generated and reference texts. For semantic evaluation, we use `bert-base-uncased` with BERTScore, computing cosine similarity between token embeddings with IDF weighting enabled and a batch size of 64. These scores are averaged across the test set to reflect global performance.

To capture narrative fluidity, we apply SegEval with a sliding window of $W = 5$ segments. The evaluation utilizes GPT-4 with predefined rubrics that cover coherence, referential consistency, and style. Each segment is compared to its surrounding reference context (oracle-based), and scoring is repeated three times to reduce variance. The module outputs aggregated metrics and error diagnostics that inform model tuning and optimization. Together, this evaluation pipeline closes the loop with the generation module, ensuring reliable performance validation across lexical, semantic, and structural levels.

For the SegEval module, GPT-4 was used in zero-shot mode with a rubric-based prompt to assess the fluency, narrative coherence, and stylistic consistency of the generated descriptions. Each segment was evaluated along three criteria: `Fluency`, `Narrative Coherence`, and `Style Consistency`, rated on a 1–5 scale. The model received the current segment along with up to five preceding reference segments, and was asked to rate the quality based on flow, logical connection, and stylistic uniformity. An example instruction was: *"Rate the following segment from 1 (poor) to 5 (excellent) based on fluency, coherence with prior context, and consistency of style."* Each segment was evaluated three times, and the scores were averaged to reduce variability. Using GPT-4 enables us to capture the subjective and qualitative aspects of descriptions, such as narrative fluidity and stylistic alignment, which are challenging to measure with classical or semantic metrics. This complements quantitative assessments, ensuring that the generated descriptions are more closely aligned with human judgment.

III. Results and Discussion

A. Impact of Parameters and Techniques on Lexical Metrics

To evaluate the influence of specific architectural and prompting choices on the lexical quality of the generated descriptions, we conducted controlled experiments varying three key factors: (i) the use of short-term memory in prompts (0 to 3 previous clips), (ii) the application of LoRA fine-tuning, and (iii) the maximum input context length (1024 vs. 2048 tokens). We report results using three standard captioning metrics: *ROUGE-L*, which measures the longest common subsequence between generated and reference texts; *CIDEr*, which evaluates n-gram similarity weighted by TF-IDF relevance; and *SPICE*, which captures semantic fidelity based on scene graph matching. All metrics are computed at the clip level and averaged across the validation set.

Figure 2 presents a grouped bar chart comparing all configurations. Incorporating short-term memory progressively improved performance. Moving from a simple prompt to one including two previous clips increased ROUGE-L from 67.4% to 73.2% and CIDEr from 1.02 to 1.26. SPICE also rose modestly, from 20.4 to 21.0. Using three clips slightly reduced performance in CIDEr (1.22), suggesting possible information redundancy or prompt overload.

TABLE I

Summary of Methodology: Techniques and Parameters by Module

Stage	**Technique**	**Parameters**
Dataset	MSR-VTT [5]	10,000 clips; standard split: 6513 train / 497 val / 2990 test
Preprocessing	CLIP-ViT-B/32	Frame sampling: every 3s; Resolution: 224×224; Output: 512-dim embeddings
	Whisper-small	Beam size: 5; Temperature: 0.7; UTF-8 token normalization
	Subtitle Alignment	Overlap threshold: 70%; Merge priority: Whisper
Generation	LLaMA 3.2–90B / Vicuna-13B	Max prompt: 2048 tokens; memory: 3 previous clips; tokenizer: `llama-tokenizer`
	LoRA Fine-Tuning	Rank $r = 8$; $\alpha = 16$; dropout $= 0.05$; 3 epochs; learning rate 3×10^{-5}; AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$
	Decoding	Greedy; Max output: 120 tokens; Temperature: 0.0; Repetition penalty: 1.2
Evaluation	ROUGE-L, CIDEr, SPICE	CIDEr $n = 4$, smoothing=3; MS-COCO eval toolkit
	BERTScore	Model: `bert-base-uncased`; IDF weighting; batch size: 64
	SegEval (GPT-4)	Context window $W = 5$; 3 repetitions per segment; metrics: fluency, coherence, style consistency

LoRA fine-tuning, applied with rank $r = 8$, $\alpha = 16$, and dropout $p = 0.05$, further improved CIDEr to 1.38 and slightly stabilized SPICE at 21.1, while ROUGE-L remained constant at 73.2%. This indicates that fine-tuning increased the model's ability to match diverse human references, especially in terms of informativeness and specificity. Additionally, prompt length had a measurable effect: limiting input to 1024 tokens reduced ROUGE-L to 71.6 and CIDEr to 1.27, while extending to 2048 tokens recovered maximum scores. These findings validate the importance of striking a balance between contextual depth and prompt manageability when generating temporally grounded narratives.

B. Semantic and Narrative Coherence Evaluation

To evaluate the deeper aspects of output quality, we measured semantic similarity and temporal cohesion using BERTScore (F1), as well as SegEval's Fluency, Coherence, and Style Consistency. As shown in Table II, performance improved across all dimensions as contextual prompting and fine-tuning techniques were incrementally applied. Notably, BERTScore increased from 86.9 with a simple prompt to 90.3 in the complete configuration using LoRA and a 2048-token context, indicating more substantial semantic alignment with reference descriptions.

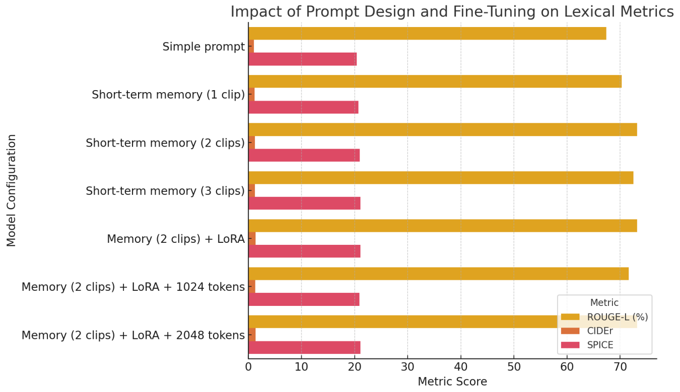


Fig. 2. Performance comparison across different configurations using ROUGE-L, CIDEr, and SPICE. Each configuration varies by the use of short-term memory (0–3 previous clips), presence or absence of LoRA fine-tuning, and prompt length (1024 vs. 2048 tokens). The chart highlights the positive impact of memory-aware prompting and LoRA on CIDEr and ROUGE-L, while SPICE remains relatively stable across settings.

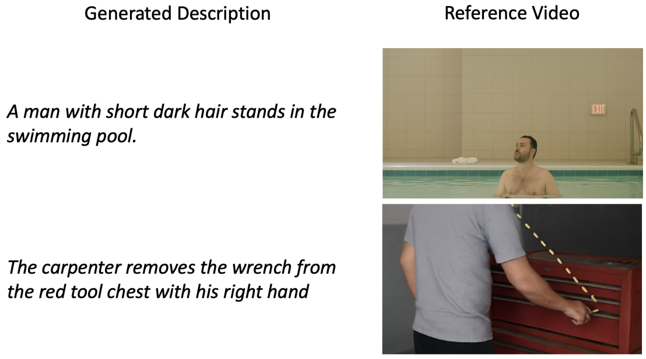

Generated Description
Reference Video
A man with short dark hair stands in the swimming pool.
The carpenter removes the wrench from the red tool chest with his right hand

Fig. 3. Examples of generated descriptions and their corresponding reference frames. These illustrate the model's ability to maintain visual-semantic alignment and descriptive consistency across segments.

Narrative structure also benefited from these enhancements. SegEval Coherence rose from 3.0 to 3.9, reflecting better referential continuity across clips. For instance, earlier configurations inconsistently alternated between expressions like “the man” and “someone,” while the optimized model maintained consistent phrasing throughout. Style consistency also improved, reaching 4.3, with more uniform sentence structures and vocabulary across segments.

Figure 3 provides concrete examples of these improvements. In the first case, the generated sentence *“A man with short dark hair stands in the swimming pool”* accurately matches the visual segment. In the second, *“The carpenter removes the wrench from the red tool chest with his right hand”* demonstrates the model's ability to ground object references in visual context, supporting the observed increase in coherence and fluency.

TABLE II
SEMANTIC AND NARRATIVE EVALUATION METRICS ACROSS CONFIGURATIONS

Configuration	BERTScore	Fluency	Coherence	Style
Simple prompt	86.9	3.3	3.0	3.5
Memory (2 clips)	88.4	3.6	3.4	3.8
Memory (2 clips) + LoRA	89.6	3.9	3.7	4.1
Memory (2 clips) + LoRA + 2048 tokens	**90.3**	**4.1**	**3.9**	**4.3**

C. Efficiency and Multimodal Coverage

In addition to accuracy and coherence, the proposed system was designed to operate efficiently in real-world scenarios. As summarized in Table III, the average generation time per clip was 1.18 seconds, making the system suitable for near real-time applications such as live captioning or assistive video summarization. Preprocessing steps—including frame extraction, transcription, and metadata alignment—took an average of 19.4 seconds per 3-minute video, allowing the system to scale with minimal latency.

From a resource standpoint, the pipeline was tested on a 24GB GPU, where it maintained a memory footprint of approximately 16.2GB during inference. The average number of tokens per prompt was 1,823, staying within the model's 2048 token limit while leveraging context effectively. This prompt size allowed for temporally rich descriptions without incurring overflow or performance degradation.

Coverage analysis revealed that the system accurately captured 82.4% of visually relevant events and 76.5% of significant auditory events compared to human references. For example, in one clip, the system accurately described a subtle head turn toward a speaker, which was visually evident but often missed by baseline models. In another example, the system successfully transcribed background laughter during a dialogue scene, improving emotional context and narrative depth. These findings highlight the system's ability to fuse visual and auditory streams into comprehensive and context-aware outputs.

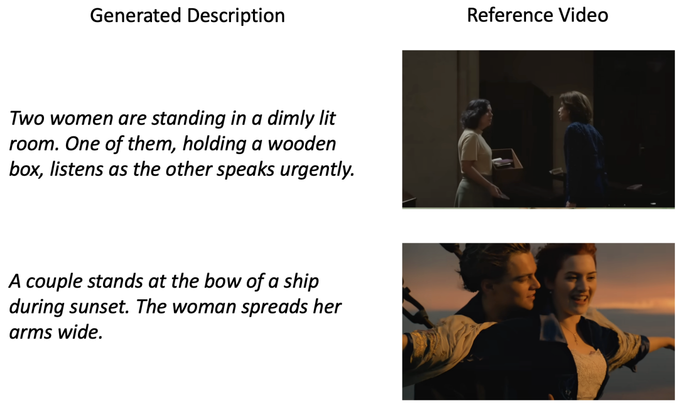
TABLE III
EFFICIENCY AND MULTIMODAL COVERAGE METRICS

Metric	Value
Average generation time per clip (s)	1.18
Average preprocessing time per 3-min video (s)	19.4
GPU memory usage (24GB GPU)	16.2 GB
Average tokens per prompt	1,823
Visual event coverage (%)	82.4
Auditory event coverage (%)	76.5

Figure 4 illustrates two representative cases in which the system demonstrates accurate multimodal alignment. In the first, it identifies an emotionally tense exchange between two women in a dark room, correctly associating the visual posture and urgency of speech with the narrative structure. In the second, it captures the iconic moment of a couple standing at the bow of a ship, describing both body language and contextual cues (sunset, motion, interaction) with temporal fluency. These examples exemplify the model's ability to synchronize vision and sound in semantically meaningful ways.

IV. REAL GAINS AND LIMITATIONS

To assess the practical benefits of the proposed system, we compared its performance to a baseline in key dimensions, as summarized in Figure 5. The results demonstrate substantial improvements: narrative coherence increased from 3.0 to 3.9, style consistency increased from 3.5 to 4.3, and fluency increased from 3.3 to 4.1, all as assessed by SegEval.



Two women are standing in a dimly lit room. One of them, holding a wooden box, listens as the other speaks urgently.

A couple stands at the bow of a ship during sunset. The woman spreads her arms wide.

Fig. 4. Examples of multimodal alignment. In both scenes, the system successfully integrates visual features (gesture, composition, expression) with auditory elements to generate semantically rich and temporally coherent descriptions.

Semantic similarity, measured by BERTScore, improved from 86.9 to 90.3, while ROUGE-L increased from 67.4 to 73.2. Furthermore, generation speed per clip decreased from 1.85 seconds to 1.18 seconds, underscoring the system's enhanced efficiency despite richer input prompting.

These gains were particularly evident in scenes that required temporal grounding or stylistic control. For example, the proposed system successfully maintained consistent character references across multiple segments, using coherent phrasing and avoiding lexical drift — a failure mode frequently observed in the baseline. It also adapted the narrative rhythm to the emotional tone, producing more fluent and engaging descriptions.

Nonetheless, some limitations persist. In cases involving overlapping dialogue and background music, Whisper occasionally failed to prioritize salient audio signals, resulting in unclear output. Similarly, in visually ambiguous scenes with minimal motion or low contrast, the model sometimes misattributed actions or omitted key details. These issues underscore the need for future research on multimodal disambiguation and improved audio source separation.

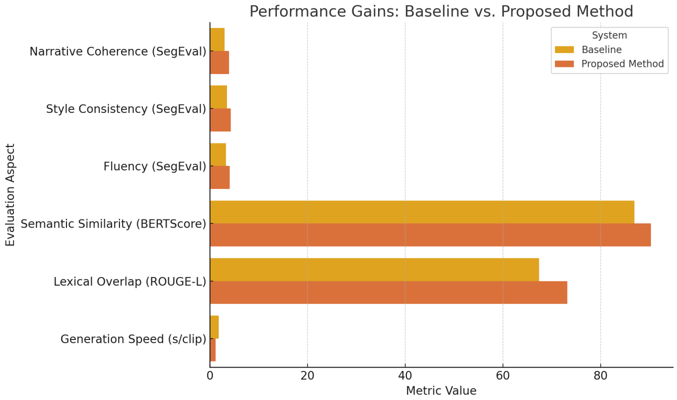


Fig. 5. Comparison of baseline and proposed method across six core evaluation metrics. The proposed approach yields consistent improvements in semantic quality, narrative structure, and generation speed.

Future work will explore integrating hybrid speech separation models, such as SepFormer (Sub-band SepFormer) or Conv-TasNet, into the Whisper transcription pipeline to address the challenges of overlapping dialogue and background noise. We plan to enhance visual encoding by incorporating optical flow and motion saliency maps into the CLIP (Contrastive Language-Image Pretraining) based embedding process for visually ambiguous scenes with low contrast or minimal motion. These strategies aim to improve the disambiguation of complex audiovisual content and increase descriptive accuracy.

V. Conclusion

This work introduced a modular and efficient system for generating coherent multimodal descriptions of long-form videos. By combining CLIP-ViT for visual embeddings, Whisper for audio transcription, and LLaMA 3.2–90B with LoRA fine-tuning for text generation, the architecture effectively integrated visual and auditory inputs into context-aware narratives. Prompt-based short-term memory ensured temporal continuity while maintaining scalability and low inference latency. Extensive evaluations confirmed consistent improvements in lexical, semantic, and narrative metrics. The system outperformed a baseline model, with gains of +3.4 in BERTScore, +0.9 in narrative coherence (SegEval), and a 36% reduction in generation time per clip. The model also demonstrated robustness in capturing subtle visual cues and emotionally charged audio events, contributing to more fluent and engaging outputs. Despite these strengths, limitations remain in scenes involving low contrast, overlapping audio, or abstract actions. Future work will address these challenges through advanced audiovisual disentanglement techniques and adaptive prompting strategies. The proposed framework also presents clear application potential in accessibility services, educational content summarization, and video indexing—highlighting its relevance beyond research, toward scalable and inclusive media technologies.

Acknowledgements

Part of the results presented in this work were sponsored by *ENVISION Indústria de Produtos Eletrônicos LTDA*, under the terms of Brazilian Federal Law nº 8.387/91 (SUFRAMA). This research was conducted with *UFAM/CETELI* and *Envision (Grupo TPV)*.

References

[1] Nayyer Aafaq, Ajmal Mian, Wei Liu, Syed Zulqarnain Gilani, and Mubarak Shah. Video description: A survey of methods, datasets, and evaluation metrics. *ACM Computing Surveys*, 52:1–37, 10 2019.

[2] Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13647–13657, Los Alamitos, CA, USA, June 2024. IEEE Computer Society.

[3] Yuhan Shen, Huiyu Wang, Xitong Yang, Matt Feiszli, Ehsan Elhamifar, Lorenzo Torresani, and Effrosyni Mavroudi. Learning to segment referred objects from narrated egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14510–14520, June 2024.

[4] Chao-Yuan Wu and Philipp Krähenbühl. Towards long-form video understanding. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, 2021.

[5] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296, 2016.

[6] Aanisha Bhattacharyya, Yaman Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. pages 9822–9839, 01 2023.