An Automatic Deep Learning Approach for Trailer Generation through Large Language Models

# An Automatic Deep Learning Approach for Trailer Generation through Large Language Models

Roberto Balestri*
*Department of the Arts*
*University of Bologna*
Bologna, Italy

Pasquale Cascarano
*Department of the Arts*
*University of Bologna*
Bologna, Italy

Mirko Degli Esposti
*Department of Physics and Astronomy*
*University of Bologna*
Bologna, Italy

Guglielmo Pescatore
*Department of the Arts*
*University of Bologna*
Bologna, Italy

***Abstract*—Trailers are short promotional videos designed to provide audiences with a glimpse of a movie. The process of creating a trailer typically involves selecting key scenes, dialogues and action sequences from the main content and editing them together in a way that effectively conveys the tone, theme and overall appeal of the movie. This often includes adding music, sound effects, visual effects and text overlays to enhance the impact of the trailer. In this paper we present a framework exploiting a comprehensive multimodal strategy for automated trailer production. Also, an Large Language Model (LLM) is adopted across various stages of the trailer creation. First, it selects main key visual sequences that are relevant to the movie's core narrative. Then, it extracts the most appealing quotes from the movie, aligning them with the trailer's narrative. Additionally, the LLM assists in creating music backgrounds and voiceovers to enrich the audience's engagement, thus contributing to make a trailer not just a summary of the movie's content but a narrative experience in itself. Results show that our framework generates trailers that are more visually appealing to viewers compared to those produced by previous state-of-the-art competitors.**

***Index Terms*—Deep Learning, Large Language Model, Multimodal Approach, Multimedia, Movie's Trailer generation.**

## I. Introduction

Movie trailers stand as important elements, delivering brief, but profound, narratives that significantly influence a film's market reception and box office success. The art of trailer creation is a synthesis of creativity and tactical planning. These trailers are more than promotional snippets: they encapsulate the essence of the films they represent, using a blend of artistry and strategic insight to captivate audiences. This creative process is intensive, demanding a deep dive into the film's narrative fabric, requiring teams to dissect and analyze scenes and dialogues to distill the film's core into a briefer narrative [1], [2].

Industry professionals employ time-consuming techniques to ensure these previews intrigue the viewers, encouraging them to explore the full movie [3]–[5]. Creating a film trailer is a highly complex task, typically requiring several days of dedicated collaboration among a skilled team. This team examines the film's content and screenplay to fully grasp its narrative structure and themes. Hence, the challenge of automating this process is immense. However, it presents an exciting opportunity to enhance trailer production, aiming to replicate the human touch in terms of aesthetic quality, rhythmic precision, narrative coherence, and emotional impact. [6]–[8].

In the Artificial Intelligence (AI) era, generative models have revolutionized various creative industries, including many filmmaking tasks [9], [10], such as script writing and story development [11], character design and animation [12], [13] and soundtrack composing [14]. These technologies, particularly Large Language Models (LLMs) [15], are reshaping traditional approaches to content creation, offering unprecedented opportunities [16]. LLMs, such as Generative Pre-trained Transformer (GPT) models [17], are at the forefront of natural language generation: they are trained on vast amounts of pre-existing textual data to produce contextually relevant (new) text. Historically, recurrent neural networks (RNNs) [18], [19] have been the main technologies for sequential data processing tasks and, in particular, for natural language processing tasks [20]. However, RNNs encountered several challenges, including vanishing gradients and difficulty in capturing long-range dependencies. As a result, LLMs became the preferred models by overcoming these limitations with transformer architectures and self-attention mechanisms [21]–[23], enabling them to understand and generate natural language with unprecedented accuracy and fluency.

So far, in the literature, there is a lack of usage of Generative AI driven systems that automatically generates movie trailers. For this reason, in this paper, we introduce a new framework for automated trailer creation. This framework utilize a LLM to orchestrate each phase of trailer production, from scene selection to the integration of on-screen and off-screen dialogues and music. The LLM's capabilities enable a full understanding of the movie plot, facilitating the assembly of a trailer that adheres to the film's narrative. Furthermore, to illustrate the capabilities of the technology, we have developed "hybrid" trailers that blend traditional voice-over techniques with the dynamic integration of actual movie dialogues. This work explores the capacity of the LLM-driven methodology to enhance automated systems with narrative insight and creativity. Our approach respects the tradition of trailer production (voice-over) while incorporating modern storytelling techniques (movie dialogues) [24].

The flow of the manuscript is organized as follows: in Section II we provide a short overview of existing works in the field of trailer generation focusing on the usage of LLM-driven

*Corresponding author: roberto.balestri2@unibo.it.

methods for video generation. Then, in Section III we report a detailed description of our framework. In Section IV we carry on a detailed analysis of our framework's outcomes by comparing them with the ones obtained by two state-of-the-art methods in the field, namely Movie2trailer [25] and PPBVAM [26]. Finally, in Section V we conclude the presentation of our paper.

The code for the framework is available at the Github repository https://github.com/robertobalestri/Trailer-Generation-Framework. Due to copyright restrictions, direct access to the generated trailers for the majority of movies cannot be provided. However, a trailer for the public domain movie *Night of the Living Dead* (1968) by George A. Romero can be viewed at https://youtu.be/O9fS8s2LRqM.

## II. Related works

The task of automatic creation of movie trailers has received less attention among the researchers if compared to the broader area of video content summarization.

Video summarization aims to condense content offering tools to manage the growing volume of video data, typically used for educational purposes, surveillance footage, and general video archives, thus not trying to craft engaging narratives [27]. This task has been largely investigated exploring different techniques [28] such as learning-based paradigms [29]–[31] and, very recently, focusing on the adoption of LLMs [32], [33].

Conversely, the automatic generation of movie trailers is an interdisciplinary research area spanning natural language processing, computer vision, and multimedia content creation aiming at the creation of an engaging and promotional snippets that attract viewers to watch the full video or movie. Prior works in this field have explored various techniques based on visual and/or auditory feature analysis. More precisely, in [26] the authors introduce a surrogate measure of video attractiveness and develop a self-correcting point process-based model to describe video attractiveness dynamics, then they propose a graph-based algorithm to generate trailers. Similarly, in [34] a graph convolutional neural network is used to extract visual and relational features of shots and selects the best ones for the trailer. A different paradigm is proposed in [35] where the authors introduce an automated method for creating movie trailers by selecting shots based on audiovisual features using a support vector machine. Analogously, in [25] the authors make use of anomaly detection strategies for shot selection. In the literature, other approaches have been developed to generate the trailer based on emotion and content analysis. In [36] an AI system is implemented integrating multi-modal semantics extraction to understand and encode emotional patterns specific to horror movies. The framework presented in [6] employs affective content analysis to identify impactful speech and video segments, which are then assembled into a trailer using an algorithm designed to maximize the emotional impact. Finally, approaches based on narrative and contextual analysis have been adopted. In [37] the authors leverage a graph-based representation of movies to identify narrative structures and predicting sentiment using screenplay text to enhance understanding of shot relationships. Similarly, in [7] natural language processing and machine learning are used to extract textual features from subtitles, to classify movies into genres, and to generate trailers accordingly.

In the field of video generation with the use of LLMs, there are notable innovations that, while not directly aimed at automatic trailer generation, highlight the potential of LLMs in the broader context of video content creation. For instance, Zhu et al. describe the MovieFactory framework that transforms textual descriptions into complete movies in [38]. This system leverages GPT models to create detailed scripts, which are subsequently brought to life using video and audio generation techniques . Similarly, in [39], the authors introduce "VideoDrafter", a framework that utilizes LLMs to convert text prompts into multi-scene scripts . While these frameworks are not specifically designed for creating movie trailers, their methodologies highlight the versatility and power of LLMs in crafting engaging narrations.

Most of the trailers generated by the aforementioned approaches lack of additional features like sound effects and music, being often a mere sequence of visual shots. Furthermore, the literature lacks of comprehensive approaches based on visual, emotional and narrative patterns all integrated in the process of generative movie trailers. Unlike existing literature, the framework proposed in this paper advances trailer generation by incorporating a Large Language Model (LLM) to coordinate the various tasks involved in the trailer creation process. Our system ensures that each component—from scene selection to soundtrack composition—contributes to a trailer that captures the narrative essence of the film.

## III. The Proposed Framework

This section presents the proposed framework. In Figure 1 we provide a conceptual map of the its architecture, which is organized into four core stages, namely Preparation Stage, Visual Stage, Voice-Over Stage and Soundtrack Stage. We point out that the LLM we used is OpenAI's GPT-4 interfaced via API calls. The framework's design is such that it could also be adapted to open-source models, suggesting that the approach is both innovative and accessible, with potential for broader application by other researchers. Hereafter, we describe the main phases and subphases of the process.

### A. Stage 1: Preparation.

This step involves an initial setup, extracting frames from the movie, and dividing the movie synopsis into scenes using the LLM.

- **Setup.** The initial setup involves setting up a few key configurations for movie trailer generation, including specifying Internet Movie DataBase (IMDB) [1] code, video file location, and project name. Movie information such as synopsis, relevant quotes, release date and director's

[1]IMBD website

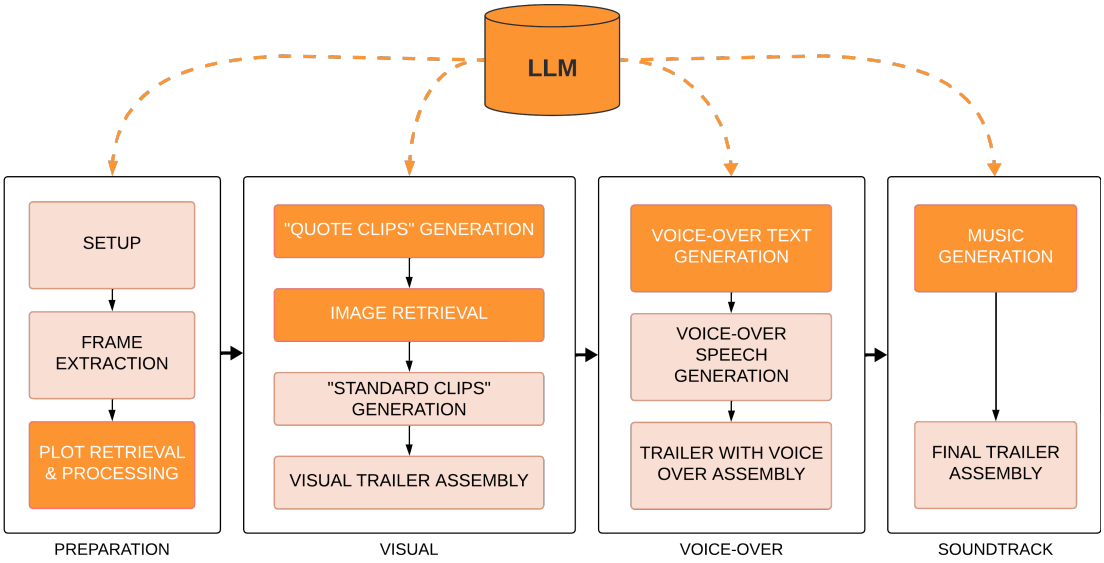


Fig. 1. The framework's structure. We report all the four principal phases and the inner subphases. The orange boxes highlight the specific subphases where the Large Language Model plays an active role.

name are scraped from IMDB using Cinemagoer [40], a Python library, and a custom script.

- **Frame Extraction.** This step focuses on extracting frames using FFmpeg [41], a Python tool that ensures a balance between frame sampling and computational costs. The frames are extracted every nine seconds (an empirically chosen number). The first and last part of the movie are not computed in order to avoid opening and ending credits frames.
- **Plot Retrieval & Processing.** This step involves refining the movie's synopsis into a plot outline using the LLM. The synopsis is first filtered locally to comply with GPT's content filters [42], then segmented into sub-plots by the LLM. Each sub-plot will represent a scene within the generated trailer, ensuring a coherent narrative arc. Sub-plots are organized for visual matching and trailer assembly. We asked the LLM to provide very visual subplots, so that we can retrieve the selected content from the movie more easily. For example, these are the first three sub-plots extracted for the movie Interstellar (2014):
  - *A dusty farm under a fading sky.*
  - *A father, Cooper, checks over crops with a knowing frown.*
  - *Young Murphy stares at mysterious patterns in the dust.*

For the sake of readability we do not report the original input prompt given to the LLM which can be found in the Github repository. However, we asked the LLM to generate clear, simple descriptions of key visual scenes for a movie trailer, focusing on straightforward language to introduce main characters and locations without revealing the movie's conclusion, prioritizing simplicity and clarity, avoiding complexity and poetic language and producing descriptive phrases.

### *B. Stage 2: Visual.*

This step involves the creation of "Quote Clips," which are segments filled with impactful dialogues (selected by the LLM from the previously scraped quotes from IMDB movie's page) that capture the essence of the film's narrative, and "Standard Clips", which are the visual backbone of the story, guiding viewers without spoken words. These elements are then assembled into a coherent visual trailer.

- **"Quote Clips" generation.** This step involves selecting key dialogues from the film to enrich the trailer's narrative depth. After initial gathering, scraped movie's quotes undergo through cleaning and filtering, including speaker separation and content standardization. Length, structural completeness, and sentiment intensity (using TextBlob [43] Python package) are checked to ensure relevance. The top 200 shortest quotes are selected, filtered from violent and sex-related words, and evaluated by the LLM for alignment with the film's themes and emotions. Audio extraction and transcription are done using StableWhisper [44], a sequence matching algorithm that aligns text quotes with audio segments, refined by Pyannote's [45] voice activity detection model. We refine the alignment with Pyannote because StableWhisper is not always precise in indicating the exact boundaries of the phrases. Video clips are extracted, and shot boundary detection (SBD) is applied [46]. Based on the results from the SBD process, the system refines clips substituting any "orphan shot" (shots that are too short to be visually appealing) at the start or at the end of the clip with a black screen to enhance visual continuity.
  For example, these are two among the selected quotes from the movie *Interstellar*:
  - *Love is the one thing we're capable of perceiving*

*that transcends time and space.*
- *We've always defined ourselves by the ability to overcome the impossible.*

More precisely, We asked the LLM to analyze movie script phrases for trailer use, identifying impactful and memorable lines evaluating their emotional and thematic impact. We omit the full prompt for the sake of brevity.

- **Image retrieval.** Inspired by a previous work from Dimitre Oliveira [47], the system aligns the film's narrative with visual representations by extracting keywords from sub-plot lines and using them as anchors. It employs the deep learning Clip-ViT-L-14 model to embed textual and visual content into a shared semantic space, ensuring semantic congruence [48], [49]. Frames exhibiting high semantic similarity to keywords (extracted from sub-plots) are selected based on temporal distribution. The selection process considers both the narrative progression and semantic alignment, quantified through cosine similarity measures. To ensure a representation of various parts of the movie, the distance in seconds between selected frames is at least 1.5% of the total duration of the movie. EasyOCR [50] and CRNN [51] are used to ensure selected frames are free of superimposed text.
  The retrieved frame for the phrase "A dusty farm under a fading sky." from *Interstellar* can be seen in Fig. 2. In this stage we asked the LLM to extract five key semantic keywords from movie plots, focusing on themes, characters, and significant events without redundancy.

Fig. 2. The frame retrieved for the sentence "A dusty farm under a fading sky." from the movie *Interstellar* .

- **Standard Clip generation.** The system creates precise video segments by establishing a "buffered zone" around selected frames to capture the full context. Shot boundary detection algorithms [46] are employed within this zone to identify start points for each clip, ensuring coherence and completeness. Like the Quote Clips generation, the system aims to prevent "orphan shots" in Standard Clips. Here, clip lengths are adjusted to improve narrative flow and ensure visual continuity.
- **Visual Trailer Assembly.** Here the algorithm combines Standard and Quote Clips to create a narrative flow. Standard Clips are sorted to reflect their original narrative order, preventing inconsistencies. Quote Clips undergo audio source separation using a hybrid transformer model [52] to retain only the vocal part, then they are dispersed among standard clips using a systematic interval strategy. For example, if a trailer has 5 clips in total, of which 2 are Quote Clips (QC) and 3 are Standard Clips (SC), we have a trailer sequence like this: SC, QC, SC, QC, SC. Attention is paid to audio transitions, implementing fade-in and fade-out effects for smoother changes. All clips are concatenated into one unified video file, and a timestamp log is generated to record the start and end times of each Quote Clip in the trailer, aiding in later audio editing.

### *C. Stage 3: Voice-Over.*

This step involves the usage of the LLM to generate a voice-over script that complements the visual content. This script is converted to audio, and the voice-over is synchronized with the visual trailer.

- **Voice-Over Text Generation.** In this phase, the LLM generates evocative phrases for the trailer's voice-over using the movie's plot summary, directorial credits, and release date. The generation ensure engagement and lyricism while avoiding spoilers. The quantity of phrases matches the trailer's length, integrating narrative and visual elements harmoniously.
  For example, these are three generated voive-over phrases for *Interstellar*:
  - *In the silence of space, hope whispers for a dying Earth.*
  - *A ghost in the dust, a code to the stars under Christopher Nolan's vision.*
  - *As hours become years, the journey for mankind's future unfolds in November.*

  We asked to the LLM to craft captivating trailer phrases that mention the director's name and release month once, mimicking a sense of anticipation before the movie hits cinemas. The original prompt given as input to the LLM can be found in the Github repository.
- **Voice-Over Speech Generation.** The algorithm converts scripted text into audio using the Coqui xtts-v2 Text-to-Speech (TTS) open-source model [53], [54]. Voice selection aligns with the movie's genre(s), employing a genre-to-voice mapping strategy for films with multiple genres. The TTS model generates audio files (that we call "Voice Clips") for each voice-over script segment. A generated voice-over phrase for the movie *Interstellar* can be found at the link https://freesound.org/people/bobe94/sounds/745217/.
- **Trailer with Voice-Over Assembly.** In this phase, voice-over audio is integrated into the trailer. Timestamps indicating Quote Clips placement are used to position Voice Clips avoiding any overlap between them. Average volume of Voice Clips informs volume adjustments for Quote Clips, ensuring seamless integration. Voice Clips

are combined with the trailer's audio track and reattached to the video.

### D. Stage 4: Soundtrack.

This step involves the usage of a music generation model which composes a unique soundtrack that aligns with the film's themes thanks to the LLM's musical direction. The final trailer is assembled by integrating this soundtrack with the voice-over and visual content, ensuring a balanced presentation.

- **Music Generation.** The LLM generates a detailed music description that aligns with the story's mood and theme. The system, employing the MusicGen text-to-audio model [55], synthesizes background music based on the LLM's description. More precisely, we asked the LLM to generate a music description inspired by the plot, focusing strictly on musical elements and instrument selection. The output of the LLM for the movie *The Lord of the Rings: The Return of the King* is reported below:

  *Instruments: Cello, dulcimer, low woodwinds, brass ensemble.*
  *Key: D minor.*
  *Tempo: Moderate to slow.*
  *Dynamics: Varied, with dynamic swells mirroring conflict.*
  *Texture: Layered, introducing one instrument at a time.*
  *Mood: Ominous with moments of somber reflection.*
  *Atmosphere: Tense, foreboding.*

  An example of the soundtrack generated for the movie *The Lord of the Rings: The Return of the King* can be found at https://freesound.org/people/bobe94/sounds/750030/.
- **Final Trailer Assembly.** In this phase, the system integrates the created music with the trailer's visuals and existing audio. An audio ducking algorithm lowers music volume during high trailer volume (e.g. during Quote Clips or Voice Clips) to preserve narrative clarity. The adjusted audio is combined with the video, including fade-in and fade-out effects for a seamless viewing experience. This phase completes the trailer, ensuring the soundtrack enhances the narrative of the trailer.

## IV. Evaluation and results

In this section, we present a comparative analysis between our framework and other leading competitors in automatic trailer generation. The list of competing methods include:

- PPBVAM - Point Process-Based Visual Attractiveness Model [26]
- Movie2trailer [25].

We compare the trailers of *The Wolverine* (2013), *The Hobbit: The Desolation of Smaug* (2013), and *300: Rise of an Empire* (2014) obtained by our system, with the outcomes produced by aforementioned approaches [2]. For consistency, all trailers were standardized to a resolution of $480 \times 360$.

[2]Competitors' generated trailers

To ensure an objective evaluation, we adopted a rigorous approach. We engaged 16 volunteers (mainly students from a class of Cinema) with varied movie tastes to review the trailers, ensuring they had not seen any of the trailers before and were unaware of the order in which the methods were applied.

TABLE I
Question items of the questionnaire submitted to the participants based on a Likert scale (1-7).

| Assessment | Question | Answer Type |
|---|---|---|
| Appropriateness | How similar this trailer looks to an actual trailer? | Likert (1,7) |
| Attractiveness | How attractive is this trailer? | Likert (1,7) |
| Appropriateness | How likely you are going to watch the original movie after watching this trailer? | Likert (1,7) |

Similarly to the evaluation carried out in [25], [26], the volunteers assessed each trailer on Appropriateness, Attractiveness, and Interest, which are metrics that reflect a trailer's ability to represent the movie, engage the audience, and pique interest. The volunteers provided ratings on a Likert [56] scale from 1 (lowest) to 7 (highest) for each of the assessments, see Table I.

We point out that, unlike PPBVAM and Movie2trailer, where the original movie soundtrack was used to replace the audio, our framework retains its uniqueness by incorporating AI generated voiceovers and soundtracks, providing a fully-automatic and comprehensive audio-visual approach. Furthermore, while we could have run our method multiple times to produce several trailers and then select the best one, the trailers generated are the outcome of a single run of the proposed software. This contrasts with our competitors, who may have chosen the best human-evaluated outputs from their frameworks.

As valuable metric, we considered the total score, namely the sum of the ratings across the three metrics. For each participant, we assumed that the method with the highest total score was considered the most effective. For example, if the trailer assessed 3 on Appropriateness, 3 on Attractiveness and 2 on Interest, its total score is 8.

In Figure 3 we depicted the number of participants who assigned to the competing methods the best score for considered movies. The barplot highlights which trailer generation method best meets the criteria set by viewers in a head-to-head comparison across different cinematic contexts. Our system consistently outperformed its competitors, suggesting its superior ability to create appealing trailers. We think that the superior quality of the soundtrack and voiceovers in our framework's trailer for *The Wolverine* explains its high ratings and significant outperformance compared to trailers generated for other movies.

Furthermore, we computed the average and median scores

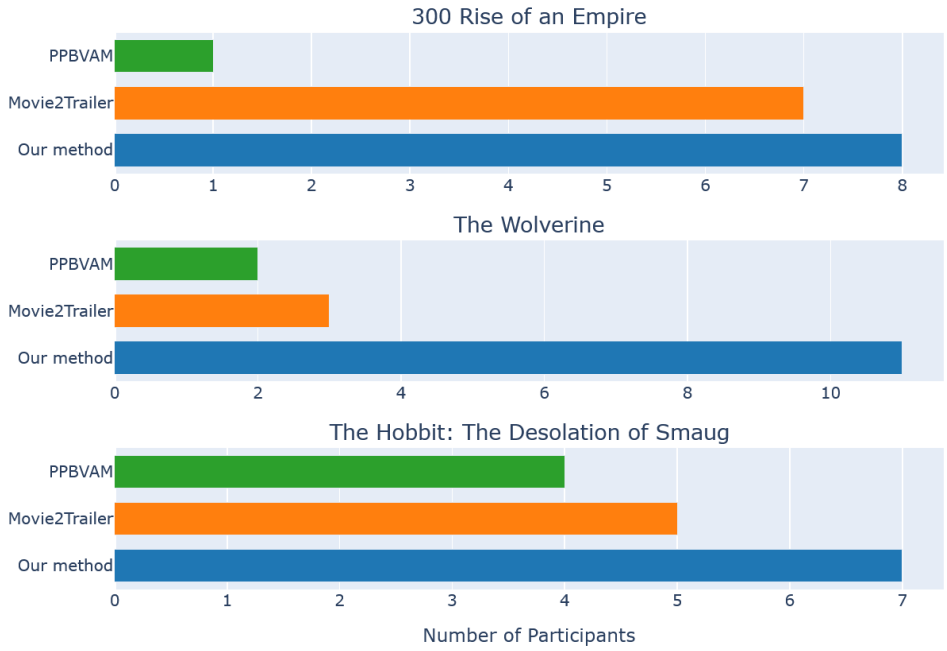


Fig. 3. Comparison of the three automatic trailer generation methods based on the total scores achieved across three movies.

for each method across the evaluated metrics for the three movies, as shown in Tables II and III, respectively. Our method demonstrated superior performance over PPBVAM in all categories and outperformed Movie2Trailer in Appropriateness and Interest. We think that the slight variance in Attractiveness is primarily due to the AI-generated soundtracks used by our system, which, despite being innovative, currently do not match the quality of the original human-composed soundtracks employed by competitors. Apart of this, our framework's performance is commendable. The success in Appropriateness demonstrates its capacity to generate trailers that align closely, compared to the competitors, with conventional expectations of how a movie trailer should look and feel. Enhancing the AI's capability to produce soundtracks that are more akin to those created by humans could further improve its standings in Attractiveness, potentially making it a leader across all evaluated categories.

TABLE II
AVERAGE SCORES BY METHOD ACROSS THE THREE CATEGORIES.

| Method | Appropriateness | Attractiveness | Interest |
|---|---|---|---|
| Our Method | **3.56** | 2.96 | **2.90** |
| Movie2Trailer | 3.15 | **3.08** | 2.85 |
| PPBVAM | 2.81 | 2.88 | 2.62 |
| | | **Mean Scores** | |

TABLE III
MEDIAN SCORES BY METHOD ACROSS THE THREE CATEGORIES.

| Method | Appropriateness | Attractiveness | Interest |
|---|---|---|---|
| Our Method | **4.0** | **3.0** | **3.0** |
| Movie2Trailer | 3.0 | **3.0** | 2.0 |
| PPBVAM | 2.0 | **3.0** | 2.5 |
| | | **Median Scores** | |

## V. LIMITATIONS, CONCLUSIONS, AND FUTURE WORK

There are few limitations in our study which are mentioned below. First, the sample size we used for the survey is pretty small, which means our findings might not apply broadly. Future research should include larger and more diverse groups to get a clearer picture on the impact of the implemented framework. Our main aim was to test if LLMs could generate trailers, not to beat the quality of human-made trailers. We went with an all-AI approach, which was innovative, but it resulted in Likert scale scores that didn't go much above average, though we did outperform existing competitors. We had to show various trailers, the survey took a while to complete. We wanted to avoid boring the participants, a factor identified by [57] as potentially biasing in studies. This made us design a survey that was a bit shallow in some areas. For instance, we didn't ask participants what elements they thought were missing in the trailers. Moreover, after reviewing our trailers, we found several areas for improvements. In future developements we will improve coherence by integrating an action recognition model to avoid including scenes in the trailer where characters speak without corresponding audio in the trailers. Indeed, we observed that in our generated trailers, Standard Clips, which are intended to be video-only, sometimes feature characters speaking, leading to disjointed and alienating results. Future works will consider the usage of novel text-to-speech models for more expressive and natural voice-overs and advanced music generation models to better match the trailer's emotional tone. Finally, we will include

more detailed quantitative and qualitative analyses to better assess our framework. Quantitatively, we'll measure how closely our generated trailers match real trailers for the same movies. Qualitatively, we'll identify specific elements our system is missing or could improve, based on a revamped survey.

In conclusion, while we're pleased with our results so far, having surpassed the state-of-the-art in automatic trailer generation, we know there's a long way to go before we can even hope to come close to the quality of human-crafted trailers.

## REFERENCES


[1] D. Hesford, "Action… suspense… emotion!': the trailer as cinematic performance," *Frames Cinema Journal*, vol. 3, 2013.

[2] D. W. Hesford, "Art of anticipation: the artistic status of the film trailer and its place in a wider cinematic culture," 2013.

[3] J. Finsterwalder, V. G. Kuppelwieser, and M. De Villiers, "The effects of film trailers on shaping consumer expectations in the entertainment industry—a qualitative analysis," *Journal of Retailing and Consumer Services*, vol. 19, no. 6, pp. 589–595, 2012.

[4] H. Krebs, "Effectful advertising? film trailers and their relevance for prospective audiences," *Telecinematic Stylistics*, 2020.

[5] R. Marich, *Marketing to moviegoers: a handbook of strategies and tactics*. SIU Press, 2013.

[6] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Automatic trailer generation," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 839–842. [Online]. Available: https://doi.org/10.1145/1873951.1874092

[7] M. Hesham, B. Hani, N. Fouad, and E. Amer, "Smart trailer: Automatic generation of movie trailer using only subtitles," in *2018 First International Workshop on Deep and Representation Learning (IWDRL)*. IEEE, 3 2018, pp. 26–30. [Online]. Available: https://ieeexplore.ieee.org/document/8358211/

[8] D. M. Argaw, M. Soldan, A. Pardo, C. Zhao, F. C. Heilbron, J. S. Chung, and B. Ghanem, "Towards automated movie trailer generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7445–7454.

[9] K. Totlani, "The evolution of generative ai: Implications for the media and film industry," *International Journal for Research in Applied Science and Engineering Technology*, 2023.

[10] J. Zhu, H. Yang, H. He, W. Wang, Z. Tuo, W.-H. Cheng, L. Gao, J. Song, and J. Fu, "Moviefactory: Automatic movie creation from text using large generative models for language and images," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9313–9319.

[11] J. J. Y. Chung, W. Kim, K. M. Yoo, H. Lee, E. Adar, and M. Chang, "Talebrush: Sketching stories with generative pretrained language models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–19.

[12] D. P. Jaiswal, S. Kumar, and Y. Badr, "Towards an artificial intelligence aided design approach: application to anime faces with generative adversarial networks," *Procedia Computer Science*, vol. 168, pp. 57–64, 2020.

[13] M. Tang, Y. Chen *et al.*, "Ai and animated character design: efficiency, creativity, interactivity," *The Frontiers of Society, Science and Technology*, vol. 6, no. 1, 2024.

[14] P. Kamath, F. Morreale, P. L. Bagaskara, Y. Wei, and S. Nanayakkara, "Sound designer-generative ai interactions: Towards designing creative support tools for professional sound designers," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–17.

[15] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier *et al.*, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and individual differences*, vol. 103, p. 102274, 2023.

[16] M. D. Esposti and G. Pescatore, "Exploring tv seriality and television studies through data-driven approaches," in *Audiovisual Data: Data-Driven Perspectives for Media Studies. 13th Media Mutations International Conference*, 2023.

[17] "Opena." [Online]. Available: https://openai.com/

[18] S. Grossberg, "Recurrent neural networks," *Scholarpedia*, vol. 8, no. 2, p. 1888, 2013.

[19] E. Loli Piccolomini, S. Gandolfi, L. Poluzzi, L. Tavasci, P. Cascarano, and A. Pascucci, "Recurrent neural networks applied to gnss time series for denoising and prediction," in *26th International Symposium on Temporal Representation and Reasoning (TIME 2019)*. Schloss-Dagstuhl-Leibniz Zentrum für Informatik, 2019.

[20] K. M. Tarwani and S. Edem, "Survey on recurrent neural network in natural language processing," *Int. J. Eng. Trends Technol*, vol. 48, no. 6, pp. 301–304, 2017.

[21] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] A. Galassi, M. Lippi, and P. Torroni, "Attention in natural language processing," *IEEE transactions on neural networks and learning systems*, vol. 32, no. 10, pp. 4291–4308, 2020.

[24] G. Richards, "Going in deep: How have movie trailers changed in the last decade?" 2018. [Online]. Available: https://www.exit6filmfestival.com/post/2018/03/14/going-in-deep-how-have-movie-trailers-changed-in-the-last-decade

[25] O. Rehusevych and T. Firman, "movie2trailer: Unsupervised trailer generation using anomaly detection," in *Proceedings of the 25th Computer Vision Winter Workshop*, D. Tabernik, A. Luke˘zi˘c, and K. Grm, Eds., 2 2020.

[26] H. Xu, Y. Zhen, and H. Zha, "Trailer generation via a point process-based visual attractiveness model," in *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-January, 2015.

[27] C. Brachmann, H. I. Chunpir, S. Gennies, B. Haller, P. Kehl, A. P. Mochtarram, D. Möhlmann, C. Schrumpf, C. Schultz, B. Stolper, B. Walther-Franks, A. Jacobs, T. Hermes, and O. Herzog, "Automatic movie trailer generation based on semantic video patterns," 2009.

[28] P. Meena, H. Kumar, and S. K. Yadav, "A review on video summarization techniques," *Engineering Applications of Artificial Intelligence*, vol. 118, p. 105667, 2023.

[29] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video summarization using deep neural networks: A survey," *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838–1863, 2021.

[30] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial lstm networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7 2017, pp. 2982–2991. [Online]. Available: http://ieeexplore.ieee.org/document/8099801/

[31] J. Xie, X. Chen, T. Zhang, Y. Zhang, S.-P. Lu, P. Cesar, and Y. Yang, "Multimodal-based and aesthetic-guided narrative video summarization," *IEEE Transactions on Multimedia*, vol. 25, pp. 4894–4908, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9797228/

[32] D. M. Argaw, S. Yoon, F. C. Heilbron, H. Deilamsalehy, T. Bui, Z. Wang, F. Dernoncourt, and J. S. Chung, "Scaling up video summarization pretraining with large language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 8332–8341.

[33] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 39–57, 2024.

[34] Y. Hu, L. Jin, and X. Jiang, "A gcn-based framework for generating trailers," in *Proceedings of the 8th International Conference on Computing and Artificial Intelligence*. ACM, 3 2022, pp. 610–617. [Online]. Available: https://dl.acm.org/doi/10.1145/3532213.3532306

[35] A. F. Smeaton, B. Lehane, N. E. O'Connor, C. Brady, and G. Craig, "Automatically selecting shots for action movie trailers," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 10 2006, pp. 231–238. [Online]. Available: https://dl.acm.org/doi/10.1145/1178677.1178709

[36] J. R. Smith, D. Joshi, B. Huet, W. Hsu, and J. Cota, "Harnessing a.i. for augmenting creativity: Application to movie trailer creation," in *Proceedings of the 25th ACM international conference on*

*Multimedia*. ACM, 10 2017, pp. 1799–1808. [Online]. Available: https://dl.acm.org/doi/10.1145/3123266.3127906
[37] P. Papalampidi, F. Keller, and M. Lapata, "Film trailer generation via task decomposition," 11 2021, arXiv:2111.08774 [cs]. [Online]. Available: http://arxiv.org/abs/2111.08774
[38] J. Zhu, H. Yang, H. He, W. Wang, Z. Tuo, W.-H. Cheng, L. Gao, J. Song, and J. Fu, "Moviefactory: Automatic movie creation from text using large generative models for language and images," in *Proceedings of the 31st ACM International Conference on Multimedia*. ACM, 10 2023, pp. 9313–9319. [Online]. Available: https://dl.acm.org/doi/10.1145/3581783.3612707
[39] F. Long, Z. Qiu, T. Yao, and T. Mei, "Videodrafter: Content-consistent multi-scene video generation with llm," 1 2024, comment: Project website: https://videodrafter.github.io arXiv:2401.01256 [cs]. [Online]. Available: http://arxiv.org/abs/2401.01256
[40] Cinemagoer, "Cinemagoer." [Online]. Available: https://cinemagoer.github.io/
[41] FFmpeg, "Ffmpeg." [Online]. Available: https://ffmpeg.org/
[42] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
[43] S. Loria, "textblob documentation," *Release 0.18.0*, 2024.
[44] jianfch, "Stable whisper." [Online]. Available: https://github.com/jianfch/stable-ts
[45] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M. P. Gill, "Pyannote.audio: Neural building blocks for speaker diarization," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, vol. 2020-May, 2020.
[46] B. Castellano, "Pyscenedetect." [Online]. Available: https://www.scenedetect.com/
[47] D. Oliveira, "Creating movie trailers with ai." [Online]. Available: TowardsAI,https://pub.towardsai.net/creating-movie-trailers-with-ai-bb5c3d89f4e3
[48] S. Transformers, "clip-vit-l-14." [Online]. Available: https://huggingface.co/sentence-transformers/clip-ViT-L-14
[49] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, 2019.
[50] JaidedAI, "Easyocr." [Online]. Available: https://github.com/JaidedAI/EasyOCR
[51] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, 2017.
[52] S. Rouard, F. Massa, and A. D'efossez, "Hybrid transformers for music source separation," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:253553270
[53] C. AI, "Xtts." [Online]. Available: https://docs.coqui.ai/en/latest/models/xtts.html
[54] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, "Libritts: A corpus derived from librispeech for text-to-speech," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2019-September, 2019.
[55] J. Copet, F. Kreuk, G. Itai, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2306.05284
[56] R. Likert, "A technique for the measurement of attitudes." *Archives of psychology*, 1932.
[57] M. Meier, C. S. Martarelli, and W. Wolff, "Is boredom a source of noise and/or a confound in behavioral science research?" vol. 11, no. 1, p. 368. [Online]. Available: https://doi.org/10.1057/s41599-024-02851-7

**DOI**: 10.1109/ICFSP62546.2024.10785516

2024 9th International Conference on Frontiers of Signal Processing (ICFSP)