Are Multimodal LLMs Movie Buffs?

Carlo Bretti, Pascal Mettes, Nanne van Noord
University of Amsterdam
{c.bretti, p.s.m.mettes, n.j.e.vannoord}@uva.nl

Abstract

No. While Multimodal Large Language Models (MLLMs) have been shown to perform very well on general video data, we systematically show that their performance on movies lags behind. This is surprising as MLLMs are increasingly used for movie understanding. To measure the performance of MLLMs on movies, we explore three pillars of movie mastery: movie knowledge, cinematographic knowledge, and critical analysis. Through a combination of quantitative and in-depth qualitative evaluations, we identify where MLLMs show promise and, in particular, where they fail. Our findings show that in small-scale settings involving factual knowledge, MLLMs are able to outperform existing methods. However, once cinematographic and critical analysis is required, MLLMs are insufficiently able to extract meaningful information from the visual modality to be able to provide useful insights. The data and project page are available at carlobretti.github.io/moviebuff.

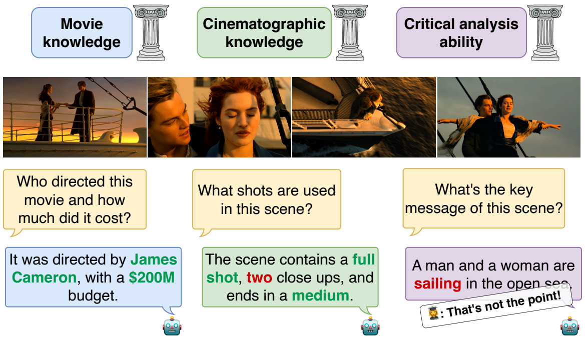

Movie knowledge
Cinematographic knowledge
Critical analysis ability
Who directed this movie and how much did it cost?
What shots are used in this scene?
What's the key message of this scene?
It was directed by James Cameron, with a $200M budget.
The scene contains a full shot, two close ups, and ends in a medium.
A man and a woman are sailing in the open sea.
: That's not the point!

Figure 1: To evaluate whether MLLMs are movie buffs, we introduce three pillars: Movie knowledge, Cinematographic knowledge, and Critical analysis. While they can perform competitively in small-scale settings, they fall short on meaningful visual movie understanding.

1 Introduction

Over the past couple of years, multimodal large language models (MLLMs) have shown great promise in video understanding tasks (Wang et al., 2024b,a; Li et al., 2024b). Among different tasks, MLLMs have been used in text-to-video retrieval (Li et al., 2024b), in video question-answering (Wang et al., 2024a), and video captioning (Wang et al., 2024b). Their strength lies in the ability to connect visual representations and to make use of LLMs' ability to understand and produce semantic predictions.

Due to the promise of MLLMs on general video understanding, recent research has begun extending their use to movies for tasks such as audio description generation (Xie et al., 2024; Han et al., 2023b,a, 2024; Ye et al., 2024; Zhang et al., 2022; Yue et al., 2023, 2025; Gao et al., 2025) or question answering (Zhang et al., 2025b; Fung et al., 2023). However, as MLLM adoption grows and MLLMs increasingly become embedded in workflows, the question remains as to how suitable they are for movie material. Since MLLMs consist of modality-specific encoders and a large language model (LLM) component, which are often trained independently, there are multiple training stages where MLLM components might come in contact with movie-related data. This variation across multiple MLLM components of how and what type of movie material they are exposed to may influence their capabilities in unique ways, which means that determining their capabilities on movies requires evaluation from multiple angles.

In this work, we investigate the extent to which off-the-shelf general-purpose MLLMs, which form the backbone of movie understanding systems, are knowledgeable about movies. Film studies handbooks often consider the understanding of stylistic and formal cinematographic elements as foundations for the ability to critically analyze a movie (Bordwell et al., 2020; Barsam and Monahan, 2019). In this spirit, we identify three pillars defining the capabilities that an MLLM should have to be considered knowledgeable about movies.

The first pillar considers cultural **movie knowl-**

Findings of the Association for Computational Linguistics: EACL 2026, pages 2661–2677
March 24-29, 2026 ©2026 Association for Computational Linguistics

edge. Movies are not created in a cultural void: they are a product of all the people and institutions involved in their creation. Among other relevant attributes, a movie involves a director, a budget, a country of production, and an intended genre. All of these elements influence the final visual product. A knowledgeable system should therefore be able to discern movies based on their cultural context.

The second capability we investigate regards classifying **cinematographic attributes**. When analyzing a movie, film scholars consider the formal visual aspects of a shot (Bordwell et al., 2020). For example, the shot size or the camera angle used are key to the look and feel of a movie. A knowledgeable MLLM should therefore be able to distinguish between different shot attributes.

Lastly, the third pillar regards the **ability to critically analyze a movie**. While film scholars are initially often trained in the formal analysis of a movie, the goal is to then be able to use formal analysis as a basis for a holistic critical analysis of a movie. In a similar vein, a capable and knowledgeable MLLM should have "taste" and be able to identify the strengths and flaws of a movie scene.

Overall, this work investigates the extent to which existing MLLMs are knowledgeable about movies. We do so by defining three pillars, and we investigate each pillar using off-the-shelf MLLMs, and by proposing two new benchmarks for movies (CMD-AD+ and MovieCom). Our results demonstrate that MLLMs show potential for movie knowledge in a small-scale setting, where movie classification is required from limited options. For classifying cinematographic attributes, which depend on visual understanding, MLLMs fall short of their specialised counterparts. Additionally, even for the critical analysis, where the textual component could potentially compensate for a lack of visual understanding, we observe a breakdown in performance, with MLLMs failing to identify key events and misinterpreting stylistic choices. Our results highlight crucial limitations in MLLMs for meaningful movie understanding.

2 Related Work

To evaluate whether MLLMs are knowledgeable about movies, we explore three pillars (cultural knowledge, cinematographic attributes, and movie critique) that lay bare their potential for movie understanding. In the following, we discuss the related work for each of these pillars.

2.1 Movie Attribute Classification

The production of movies involves a range of roles and factors that influence the final product, such as the director, where the movie was made, or how much it cost to make. These factors influence the interpretation of a movie, i.e., a viewer may be more forgiving of bad visual effects in a low-budget indie movie than they would be for a blockbuster production. One's familiarity with and knowledge about movies is reflected in the ability to recall or determine these factors or attributes of a movie. Hence, to evaluate the movie knowledge of MLLM, we focus on their ability to classify movie attributes. Over the years, many works have investigated how to classify attributes of movies (Rasheed and Shah, 2002; Zhou et al., 2010; Simões et al., 2016; Wu and Krahenbuhl, 2021).

Recently, we have observed a resurgence of work on movie attribute classification, partly driven by the release of the Long-form Video Understanding (LVU) benchmark (Wu and Krahenbuhl, 2021). Common approaches applied to the LVU benchmark include state-space modelling (Islam and Bertasius, 2022; Wang et al., 2023; Li et al., 2024a), contrastive learning (Xiao et al., 2022; Chen et al., 2023), and, of particular interest for this paper, MLLMs (He et al., 2024; Azad et al., 2025). MA-LLM (He et al., 2024) proposes a technique to improve MLLM for handling long context, which is particularly relevant for dealing with the long temporal dependencies in movies. More recently, HierarQ (Azad et al., 2025) proposed a hierarchical approach that employs memory banks across two time-scales. Following a similar supervised setting, they outperform MA-LLM on movie understanding tasks. While adding additional components, both MA-LLM (He et al., 2024) and HierarQ (Azad et al., 2025) rely on pre-trained and frozen MLLMs, underscoring the importance of general-purpose MLLMs as building blocks for movie understanding systems. Therefore, we focus on general-purpose MLLMs and evaluate them on movie understanding tasks in a zero-shot manner. To further facilitate such analysis, we propose a new movie attribute dataset that enables a broader exploration of the importance of domain knowledge for attribute classification.

2.2 Cinematography Classification

The second pillar focuses on a cinematographic understanding of individual shots from a formal perspective. A common goal here is to classify shot size (sometimes called shot type or scale), which regards the scale of human bodies in the shot (Bordwell et al., 2020), such as close-ups or wide shots. Increasingly, additional cinematographic aspects such as camera motion (Rao et al., 2020; Lin et al., 2025), camera angle (Argaw et al., 2022) and lighting (Liu et al., 2025) are also considered.

The use of additional cinematographic shot attributes is also reflected in datasets. While CineScale (Savardi et al., 2021) consists of frames of movies from 6 directors annotated for shot scale, CineScale2 (Savardi et al., 2023) expands this with annotations for camera angle and level. MovieShots (Rao et al., 2020) similarly contains 47k annotated movie shots for shot scale and camera movement. More recently, the CameraBench (Lin et al., 2025) dataset was introduced, which provides extremely detailed annotations for the type of camera movement employed. However, particularly notable is AVE (Argaw et al., 2022), which is the largest available dataset, containing over 200k annotated videos with annotations for multiple attributes, including shot size, shot type, camera motion, and camera angle. Approaches for AVE mainly feature large-scale multimodal contrastive pretraining (Argaw et al., 2023) and multi-task training, trying to optimize for the commonality between the different tasks at hand (Argaw et al., 2022; Li et al., 2024c). More recently, two additional datasets have been introduced, ShotBench (Liu et al., 2025) and CineTechBench (Wang et al., 2025), which cover a similar range of attributes as AVE. However, given AVE's larger size and scope, we focus on AVE to get a more accurate judgment of MLLMs' cinematographic understanding abilities.

2.3 LLMs for Analysis and Critique

While the first two pillars focus on assessing knowledge, the third pillar focuses on MLLMs' ability to critically evaluate a movie scene. Recently, the Movie Facts and Fibs dataset (Zaranis et al., 2025) was introduced to evaluate the narrative understanding abilities of MLLMs' on films, by testing whether they can reasonably judge the veracity of a series of claims. In contrast, our focus is on evaluating the ability to produce free-form analyses of a movie scene. On the other hand, the ability for MLLMs to criticize and analyze visual material such as photographs (Qi et al., 2025) has been investigated. And while MLLMs have not been broadly used for visual analysis of movies, their text-only counterparts have been used for analyzing movie reviews (Paiva and Diecke, 2024). In general, LLMs have been investigated for their potential to aid in annotation for social science research (Ziems et al., 2024; Karjus, 2025). As an example, LLMs have been used to discover narratives used in discourse surrounding veganism in YouTube comments (Pera and Aiello, 2024). Recently, some works have also tested out LLMs' ability to act as academic reviewers (Tyser et al., 2024). Within the movie domain, our goal is to explore MLLMs' ability to engage in meaningful discourse about movie scenes.

3 Do MLLMs Know Who Directed Jaws?

Pillar 1. To test MLLMs' understanding of movie knowledge, we perform classification on an existing movie attribute dataset, LVU (Wu and Krahenbuhl, 2021), and on the larger dataset CMD-AD (Han et al., 2024), which we expand with movie attributes (resulting in CMD-AD+). To evaluate the performance of the MLLMs on these tasks, we prompt them in a zero-shot multiple-choice question answering setting. The aim of this evaluation is to determine what knowledge MLLMs have about movies, and to learn which information in the prompt influences their ability to correctly classify movies. Here, movie knowledge concerns both what an MLLM can infer about a movie from the prompt, and what it *knows* given its training data, i.e., we would expect a movie buff MLLM to recognize a Wes Anderson movie from his style, but similarly it should *know* that James Cameron directed Titanic from the title.

To determine the influence of additional information, we prompt the model in different ways. In the minimalistic setting, the model is prompted with either only a video clip or only textual information. We build on these settings by prompting the model with additional contextual knowledge about the movie in the prompt, such as the title, cast, and information about all other attributes that are not the one being classified. This setup allows us to compare the relative importance of data modalities.

In the following, we describe the experimental setup, data, and comparison with previous work on movie attribute classification.

3.1 Experimental Setup

For evaluation we prompt 4 MLLMs, Qwen2.5-VL (Bai et al., 2025), VideoLLaMA3 (Zhang et al., 2025a), InternVL3 (Zhu et al., 2025), and Gemma3n (Google, 2025). Using the 7B versions of both Qwen2.5 and VideoLLaMA3, and the 8B versions of InternVL3 and Gemma3n (with 4B active parameters at a time). We feed videos at one frame per second to the models following existing work (Zhang et al., 2025a), up to a maximum of 180 frames (for Gemma3n this is capped at 30 frames due to compute cost).

The task is set up as a randomized multiple-choice QA, in line with Pezeshkpour and Hruschka (2024), with the additional context in the question. The full prompt can be found in Appendix A.1.

3.2 Data

For the experiments in this section, we use two datasets, LVU and CMD-AD+. LVU (Wu and Krahenbuhl, 2021) contains video clips from 2797 movies from the MovieClips YouTube channel, split across 9 different tasks, including director, genre, writer, and year classification. For each task, LVU contains a subset of datapoints matching a small number of labels. We obtain all available videos for the existing test split to compare our results with previous work. While LVU provides a strong basis for attribute classification, the range of classes covered for each attribute is limited in scope.

In addition, we collect the 8687 available YouTube videos used in CMD-AD (Han et al., 2024) and their respective attribute information from IMDB. This newly proposed dataset (CMD-AD+) is the largest available movie attribute dataset with both attributes and video. In terms of attributes, we retrieve, on top of director and genre, labels for budget and primary country of production. For this dataset, differently from LVU, which uses 3 unique combinations of genre as class, we treat genre as a multi-label classification problem with 22 possible classes, and 439 unique combinations of genres. For budget classification, we exclude all movies without budget information. To adjust for inflation and currency, we compute the inflation-adjusted budget in USD in 2025 and bin it into 5 categories. A full comparison of LVU and CMD-AD+ can be found in Appendix A.2.

3.3 Results

To evaluate MLLMs on movie attribute classification, we experiment with four MLLMs across two datasets (LVU and CMD-AD+) and with contextual knowledge in the text, such as title, cast, or other attributes not currently evaluated. Table 1 shows the results for LVU compared to baselines (VideoBERT (Sun et al., 2019), Object-Transformer (Wu and Krahenbuhl, 2021), Orthoformer (Patrick et al., 2021), VIS4mer (Islam and Bertasius, 2022), TranS4mer (Islam et al., 2023), S5 (Wang et al., 2023), Movies2Scene (Chen et al., 2023), VideoMamba (Li et al., 2024a), MA-LMM (He et al., 2024), HierarQ (Azad et al., 2025)). A breakdown of per-attribute performance for increasing amounts of contextual knowledge can be found in Appendix A.4. For LVU, we observe that MLLMs only outperform the baseline models when relying on both visual and textual data, with just Qwen2.5-VL performing competitively in the video-only setting. However, InternVL3 surprisingly achieves the highest performance overall using just text, even surpassing the most recent video baselines. This indicates that the

	Video	Text	Avg.
VideoBERT	✓		43.3
Obj_T4mer	✓		43.6
Orthoformer	✓		50.3
VIS4mer	✓		52.7
TranS4mer	✓		53.1
S5	✓		58.0
Movies2Scene	✓		59.6
VideoMamba	✓		58.4
MA-LMM	✓		64.5
HierarQ	✓		70.0
Qwen2.5-VL		✓	68.0
	✓		68.0
	✓	✓	74.1
VideoLLaMA3		✓	61.7
	✓		57.1
	✓	✓	72.0
InternVL3		✓	77.1
	✓		46.7
	✓	✓	76.0
Gemma3n		✓	60.4
	✓		49.0
	✓	✓	65.7

Table 1: Average classification accuracy on the LVU dataset on director, genre, writer, and year. The top half includes prior works that were evaluated only on video data, the lower half compares four general-purpose MLLMs on text-only, video-only, and multimodal.

	Video	Text	Budget	Country	Director	Genre	Avg.
Qwen 2.5-VL		✓	39.7	25.0	37.3	60.2	40.5
	✓		24.4 ± 2.3	15.9 ± 1.4	22.6 ± 0.4	45.9 ± 0.5	27.2
	✓	✓	40.2 ± 2.9	28.2 ± 1.6	38.6 ± 0.4	59.9 ± 0.3	41.7
Video LLaMA3		✓	38.3	27.6	26.9	51.4	36.1
	✓		28.4 ± 1.3	8.9 ± 0.6	10.7 ± 0.6	35.5 ± 0.4	20.9
	✓	✓	38.1 ± 1.7	31.1 ± 2.4	27.6 ± 1.0	48.0 ± 1.0	36.2
InternVL3		✓	27.5	24.6	43.2	66.2	40.4
	✓		27.6 ± 1.1	8.8 ± 0.9	19.9 ± 0.4	43.0 ± 0.4	24.8
	✓	✓	34.7 ± 1.5	31.3 ± 3.1	44.2 ± 0.8	62.4 ± 0.9	43.1
Gemma3n		✓	33.0	33.9	40.9	61.3	42.3
	✓		22.0 ± 0.7	12.0 ± 0.8	16.3 ± 0.4	46.4 ± 0.4	24.2
	✓	✓	26.8 ± 0.5	34.7 ± 4.8	32.5 ± 1.1	63.0 ± 0.4	39.2

Table 2: Movie attribute classification F1-score on the CMD-AD+ dataset. CMD-AD+ contains multiple video clips per movie. As such, we report the average (and standard deviation) of randomly sampling a single video clip per movie 5 times. As the text-only settings do not consider the video clips, these are based on a single run.

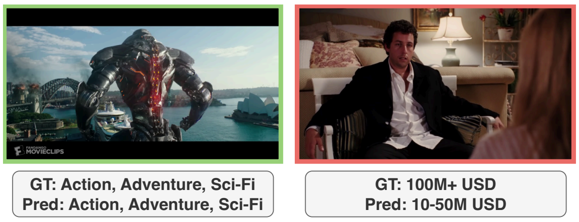

GT: Action, Adventure, Sci-Fi
Pred: Action, Adventure, Sci-Fi
GT: 100M+ USD
Pred: 10-50M USD

Figure 2: Qualitative results for Qwen2.5 on video only. Some movie attributes are more easily identifiable based on visuals, while for attributes such as budget, external factors, such as casting, can play a pivotal role.

text-based LLM component plays a significant role in correctly classifying movie attributes.

Our evaluation shows that the combination of visual and textual information generally boosts the performance. This shows promise in how the addition of contextual information can significantly boost the performance of off-the-shelf general-purpose MLLMs, outperforming existing MLLM-based methods such as MA-LMM (He et al., 2024) and HierarQ (Azad et al., 2025), which were specifically tailored and trained for movie understanding.

To evaluate MLLMs in a more challenging setting, we report results for movie attribute classification on the proposed CMD-AD+ dataset in Table 2. Across the board, we observe that CMD-AD+ is much more challenging than LVU, as it contains more movies and more classes per attribute. Similar to the prior results, we can observe that the text-only setting performs best across all models, with reduced performance in the visual-only setting. Although we see the highest performance for Qwen2.5-VL and InternVL3 in the multimodal setting, the gap remains small relative to the amount of additional data and processing employed.

Across the different attributes, we can see differences in performance, in particular for genre and country, which have a comparable number of classes. We suspect that genre may be discussed more directly in discourse on movies in the training data, whereas the production country may be discussed more indirectly, e.g., by mentioning movie studios or cities instead. Admittedly, classifying movie attributes such as genre or budget is not an easy feat. As shown in Figure 2, visually classifying some genres might be easier than others, e.g., a giant alien robot walking towards a city is a clear indicator of a sci-fi movie. However, for an attribute like budget, there is a strong reliance on context and domain knowledge: the high budget of the scene on the right may be better explained by its casting of the popular actor Adam Sandler than by visual aesthetics. As such, multimodal approaches should have an advantage in this domain. We further explore the relationship between popularity and task performance using box office as a proxy in Appendix A.3.

Overall, we observe that on LVU the off-the-shelf MLLMs achieve state-of-the-art performance, often outperforming prior supervised approaches. However, LVU is rather small-scale, as shown in Table 5, and only has a few classes for each

		Angle	Motion	Size	Type	Avg.
Trained on AVE						
AVE (w/o logit adj.) (Argaw et al., 2022)		28.9	31.2	39.1	62.3	40.4
AVE (Argaw et al., 2022)		49.8	43.7	67.6	66.7	57.0
LMP (Argaw et al., 2023)		57.7	46.1	67.4	63.8	58.8
IAI-AVE (Li et al., 2024c)		85.7	48.3	71.4	51.6	64.3
Zero/few-shot						
Qwen2.5-VL	ZS	45.8	28.8	55.7	53.6	46.0
	FS	52.4	31.2	60.2	50.7	48.6
	PE	47.5	27.1	55.8	58.7	47.3
VideoLLaMA3	ZS	48.6	30.5	50.2	58.6	47.0
	FS	40.6	22.0	51.7	35.2	37.4
	PE	47.3	29.8	51.8	57.2	46.5
InternVL3	ZS	45.1	28.8	55.7	53.4	45.8
	FS	54.4	29.6	60.2	56.4	50.1
	PE	52.6	27.8	54.8	62.2	49.3
Gemma3n	ZS	45.5	25.1	48.3	49.6	42.1
	FS	30.6	22.2	40.7	27.2	30.2
	PE	47.0	25.3	43.9	50.9	41.8

Table 3: Shot attribute classification mean-per-class accuracy on the AVE dataset. Prior work has been trained on the AVE dataset (including using specific class imbalance mitigation strategies), whereas the MLLMs are evaluated in a zero-shot (ZS), few-shot (FS), or prompt engineering (PE) setting.

attribute. The performance on the larger-scale CMD-AD+ is more indicative of the real movie knowledge capabilities, as it is a more challenging dataset than LVU. By scaling up to more movies and more classes per attribute, the tasks become more challenging, which is reflected in severely reduced MLLM performance. Notably, on CMD-AD+, the performance with just visual information is much worse, with the MLLM relying on the textual modality for its performance.

4 Do MLLMs Know What a Close-up Looks Like?

Pillar 2. Second, we focus on testing MLLMs' knowledge of cinematography. We focus here on the classification of cinematographic attributes of movie shots and compare our findings with previous work on the AVE dataset (Argaw et al., 2022). We test the MLLMs in three settings: a zero-shot, a few-shot with three example frames per class (one for Gemma3n due to compute cost), and prompt engineering with class definitions in text.

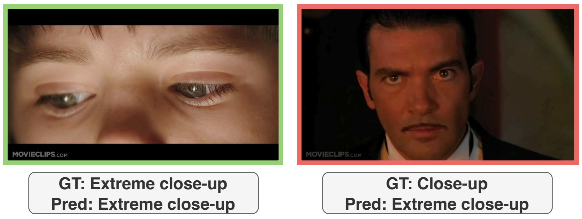


Figure 3: Qualitative examples of shot size classification by VideoLLaMA3.

4.1 Data

The AVE (Argaw et al., 2022) dataset contains 5591 movie clips with around 200k annotated shots for different characteristics. We focus on shot type, angle, motion, and size, as these are more prominent in existing movie literature (Bordwell et al., 2020) when discussing cinematographic attributes.

Most shots are very brief, with over a third of the test set being under one second. For that reason, for AVE we increase the number of frames per second processed from 1 to 4, to ensure that we sample multiple frames per clip. We split the data following (Argaw et al., 2022) and randomly sample the same number of clips for the test set.

4.2 Results

The results on AVE are shown in Table 3. Although VideoLLaMa3 performs best in the zero-shot setting, the best overall results are obtained by InternVL3 in the few-shot setting. The benefits of prompt engineering and few-shot examples appear to be mixed, with sharp drops in performance for some models. Yet, on the whole, the MLLMs are outperformed by prior approaches on all tasks, even with the additional data and compute of the prompt engineering and few-shot examples. As the data in AVE is heavily imbalanced, previous approaches employ a class imbalance mitigation strategy such as logit adjustment (Menon et al., 2021) or downsampling the training data based on prior knowledge from film production (Li et al., 2024c). As the MLLMs are evaluated without further training, similar strategies cannot be employed, nevertheless given that the MLLMs employ much larger models which were trained on large-scale datasets, we would expect more.

The MLLMs seem to have limited out-of-the-box knowledge of cinematographic concepts and lack the ability to consistently recognize them.

Among the attributes, prior research has predominantly focused on shot size, and it is frequently brought up in analyses of cinematography. To dive deeper into the results, in Figure 3 we show qualitative examples of predictions made by VideoLLaMA3 given its superior performance in the zero-shot setting. On the left, the shot is correctly classified as an extreme close-up. On the right, we see a case where the model misclassified a close-up as an extreme close-up. While a misfire, the shot is indeed a case in which the lines between the two types of shots are more subtle.

While the MLLMs are outperformed by prior work, they do demonstrate a rough understanding of the cinematographic concepts discussed and are able to recognize them visually in certain cases. In particular, we observe that the MLLMs struggle with motion, despite being given more frames (4 FPS) than for the movie knowledge tasks (1 FPS). Presumably, this attribute would benefit from video as input, as it is also challenging for prior methods, which are similarly frame-based. However, for the time being, we conclude that the cinematographic knowledge of MLLMs is lacking.

5 Can MLLMs Critique a Film Scene?

Pillar 3. The final pillar we investigate concerns MLLMs' ability to critically analyze a movie scene. To do so, we collect a new dataset of movie clips with director commentary and prompt the MLLMs to produce descriptions of the scene as if they were the director. We choose to focus on director commentary as it best captures the intent and meaning behind a scene. We then measure hallucination and omission rates (Rawal et al., 2025) across models. In addition, for qualitative assessment, we asked a film scholar to judge generated movie scene analyses as if they were students' assignments.

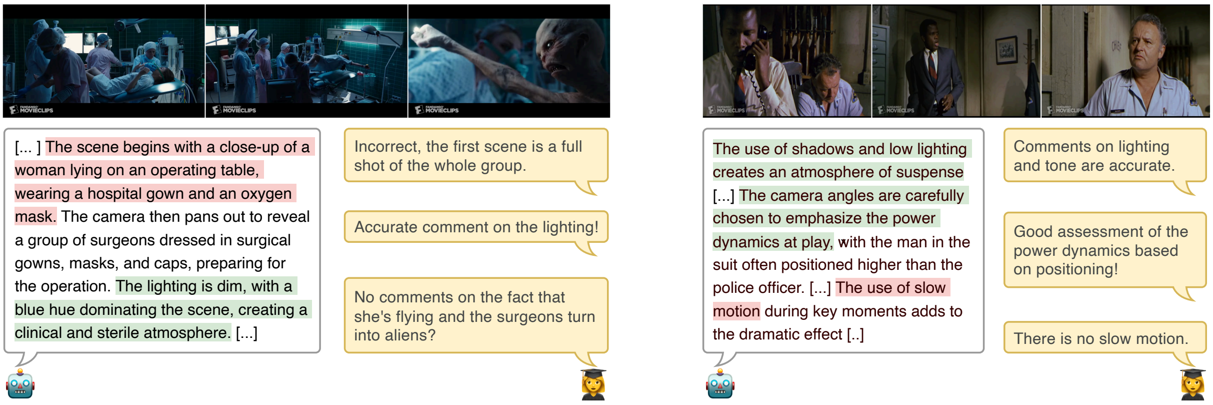
5.1 Setup

We collect a new dataset of 321 movie clips with director commentary from the Anatomy of a Scene playlist, which we dub the MovieCom dataset. We transcribe and diarize the dataset using WhisperX (Bain et al., 2023; Radford et al., 2022), and remove all speech that is not produced by the director. We then prompt the MLLMs to act as if they were the director and describe a scene and its core message. To evaluate the generated descriptions, we follow the Argus framework (Rawal et al., 2025) and use an LLM to judge the output, in line with other works on evaluating hallucination rates for MLLMs (Jing et al., 2024; Sun et al., 2024). We compute two measures: ArgusCost-H, quantifying the rate of hallucinated statements, and ArgusCost-O, which quantifies the rate at which important details are omitted. For the LLM judge, we use the 70B variant of Llama 3.3 (Grattafiori et al., 2024).

In addition, to gain deeper qualitative insights into the MLLM outputs, we prompted VideoLLaMA3 to obtain a critical analysis of a scene covering formal aspects of the movie clip and asked a film scholar to review it. For this, we randomly select 7 clips from CMD-AD+ and pass them to VideoLLaMA3 with a prompt that can be found in Appendix A.1.

5.2 Results

Quantitative results Table 4 shows the results of evaluating the descriptions produced by the MLLMs. What stands out is that all models produce high omission rates, effectively failing to describe the core message of the movie clips. While the director has a deeper understanding of the movie, the clips shown are focused on the core message, which is necessary for meaningful movie


[...] The scene begins with a close-up of a woman lying on an operating table, wearing a hospital gown and an oxygen mask. The camera then pans out to reveal a group of surgeons dressed in surgical gowns, masks, and caps, preparing for the operation. The lighting is dim, with a blue hue dominating the scene, creating a clinical and sterile atmosphere. [...]

Incorrect, the first scene is a full shot of the whole group.

Accurate comment on the lighting!

No comments on the fact that she's flying and the surgeons turn into aliens?

The use of shadows and low lighting creates an atmosphere of suspense [...] The camera angles are carefully chosen to emphasize the power dynamics at play, with the man in the suit often positioned higher than the police officer. [...] The use of slow motion during key moments adds to the dramatic effect [..]

Comments on lighting and tone are accurate.

Good assessment of the power dynamics based on positioning!

There is no slow motion.

Figure 4: Qualitative results. Here we show the results prompting an MLLM to get a critical formal analysis, together with comments from a film scholar.

understanding. In turn, we similarly observe high hallucination rates. We suspect that the textual component of the MLLMs is trying to compensate for the limited visual understanding by hallucinating a response. Upon closer inspection of the omission and hallucination rates split by statement type, we find that statements about visual descriptions have the highest rates of hallucination, while for omissions the type of statement does not affect it. Further discussion of the results by statement type are include in Appendix A.8. In general, despite differences in model architectures, there appear to be no notable differences in performance.

	Hallucination	Omission
Qwen2.5-VL	83.4	91.6
VideoLLaMA3	89.7	96.5
InternVL3	82.0	95.5
Gemma3n	86.6	93.7

Table 4: Hallucination (ArgusCost-H) and Omission (ArgusCost-O) cost of different MLLMs on the MovieCom dataset. All MLLMs hallucinate statements at a high rate and omit important statements present in the original captions.

Qualitative analysis In Figure 4 we show excerpts of the analyses made by the MLLM, including comments provided by the film scholar who reviewed them. These excerpts capture patterns in the MLLM analyses. For instance, the left example showcases how the model can pick up on the lighting used, but misconstructs what the first shot looks like, by mixing up a close-up and a full shot. Another interesting aspect is that the scene shown involves sci-fi elements, with people floating and surgeons turning into aliens, but these supernatural aspects are completely missing from the analysis. This may be due to a lack of exposure to sci-fi data in the training of the visual encoder.

On the right, the comments on lighting and tone were deemed accurate by the film scholar. In addition, the analysis contained a good assessment of the power dynamics between the two characters based on the shot composition. However, the comments on editing were incorrect, as slow motion is not employed in any of the scenes.

Overall, MLLMs can provide scene descriptions, but they are unable to engage in meaningful discourse on film. They can often correctly describe the setting of a scene and the lighting used. However, their comments are very generic, and the more specific comments are mainly incorrect and hallucinatory. On the whole, the analyses considered fail to capture key elements of a scene and the events that transpire.

6 Conclusion

In this paper, we investigated MLLMs and the extent to which they can be considered movie buffs. For this, we defined three pillars: Movie Knowledge, Cinematographic knowledge, and Critical analysis ability, and evaluated how well MLLMs perform on tasks within these pillars.

Movie knowledge. We find that MLLMs can outperform existing approaches in a small-scale setting, but that they fall short on the larger-scale setting of our newly proposed dataset CMD-AD+.

Cinematographic knowledge. MLLMs underperform compared to existing supervised approaches, highlighting their limited visual capabilities for movie understanding.

Critical analysis ability. MLLMs can identify some themes in movie clips, but miss out on key events and misinterpret stylistic choices, leading to high degrees of hallucinations.

Overall, we conclude that while MLLMs show potential for general video understanding, they lack the ability to meaningfully understand a movie scene. They are unable to recognize key plot points or crucial cinematographic qualities that may enable further analysis. Moreover, we observe that much of the MLLM capabilities heavily depend on the textual components, with the visual side severely lagging behind. While powerful and generalizable, the representations obtained from the visual encoder are often not fine-grained enough to enable the deep understanding needed for movie scenes. Future work should focus on improving the capabilities of the visual encoder to match those of the text component.

7 Limitations

In testing out the capabilities of MLLMs for movies, one key limitation is the difficulty in disentangling how each modality influences performance. This is especially apparent for Pillar 1 when discussing the influence of adding contextual information about the movie in the text prompt. While we see a clear benefit in the extra context information, it is unclear what brings the largest improvements.

The exploration of few-shot examples and prompt engineering is limited to Pillar 2. This is for two reasons: (1) the cinematographic attributes considered can be represented by a single frame, making it feasible to fit them in the context of the MLLMs considered, (2) they represent more concrete concepts that can be easily described in text. Ideally, these settings could also be explored for Pillar 1, but it is currently not feasible to fit multiple videos in the model context or to accurately describe concepts relating to movie domain knowledge.

Another limitation pertains to the subjectivity of the task introduced alongside the MovieCom dataset in Pillar 3. While the director's commentary provides a clear signal as to what the key components of a scene are, there is no objective gold standard for a scene description. Cultural products are subject to personal interpretation and can have different meanings for different individuals. Still, unfiltered commentary from the main person responsible for the creation of a film is a useful proxy for a high quality description. In a similar vein, having a single expert evaluator in Pillar 3 could be considered a weakness. However, the person's expertise as a film scholar and lecturer makes their judgement particularly valuable, especially because the task is hermeneutic in nature, rather than a subjective scoring task. In hermeneutic evaluation, the validity comes from the evaluator's domain knowledge, not from multiple aggregated independent evaluations (Moss, 1994).

8 Ethical Considerations

Given the ongoing debate on the use of AI within the film industry, we wanted to acknowledge our awareness of these issues. We aim to support better film research through a deeper understanding of the performance of MLLMs on movies and how this may lead to better analyses of movies. Although using MLLMs for such research may have consequences for data annotation workers, the current trends in research funding in the Humanities are negatively impacting possibilities for data annotation, which may make the use of MLLMs a necessity for large-scale analysis.

Additionally, we have strived to handle the difficulties around working with movie data that has complicated copyright circumstances to the best of our ability. While MLLMs have most likely been trained on copyrighted data, we believe that the cultural importance of movies makes it crucial to evaluate how these models perform on them. The datasets proposed in this paper are intended for research purposes only.

References

Dawit Mureja Argaw, Fabian Caba Heilbron, Joon-Young Lee, Markus Woodson, and In So Kweon. 2022. The Anatomy of Video Editing: A Dataset and Benchmark Suite for AI-Assisted Video Editing. In *ECCV*.

Dawit Mureja Argaw, Joon-Young Lee, Markus Woodson, In So Kweon, and Fabian Caba Heilbron. 2023. Long-range Multimodal Pretraining for Movie Understanding. In *ICCV*.

Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. 2025. HierarQ: Task-Aware Hierarchical Q-Former for Enhanced Video Understanding. In *CVPR*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-VL Technical Report. *Preprint*, arXiv:2502.13923.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proc. Interspeech 2023*.

Richard Meran Barsam and Dave Monahan. 2019. *Looking at Movies: An Introduction to Film*.

David Bordwell, Kristin Thompson, and Jeff Smith. 2020. *Film Art: An Introduction*.

Shixing Chen, Xiang Hao, Xiaohan Nie, and Raffay Hamid. 2023. Movies2Scenes: Learning Scene Representations Using Movie Similarities. In *CVPR*.

Yi Fung, Han Wang, Tong Wang, Ali Kebarighotbi, Mohit Bansal, Heng Ji, and Prem Natarajan. 2023. DeepMaven: Deep Question Answering on Long-Distance Movie/TV Show Videos with Multimedia Knowledge Extraction and Synthesis. In *EACL*.

Yingqiang Gao, Lukas Fischer, Alexa Lintner, and Sarah Ebling. 2025. Audio Description Generation in the Era of LLMs and VLMs: A Review of Transferable Generative AI Technologies. In *NAACL Findings*.

Google. 2025. Gemma 3n model overview. https://ai.google.dev/gemma/docs/gemma-3n.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023a. AutoAD II: The Sequel – Who, When, and What in Movie Audio Description. In *ICCV*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023b. AutoAD: Movie Description in Context. In *CVPR*.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD III: The Prequel - Back to the Pixels. In *CVPR*.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *CVPR*.

Md Mohaiminul Islam and Gedas Bertasius. 2022. Long Movie Clip Classification with State-Space Video Models. In *ECCV*.

Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, and Gedas Bertasius. 2023. Efficient Movie Scene Detection using State-Space Transformers. In *CVPR*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. FaithScore: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models. In *EMNLP Findings*.

Andres Karjus. 2025. Machine-assisted quantitizing designs: Augmenting humanities and social sciences with artificial intelligence. *Humanities and Social Sciences Communications*.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. 2024a. VideoMamba: State Space Model for Efficient Video Understanding. In *ECCV*.

Yili Li, Jing Yu, Keke Gai, Bang Liu, Gang Xiong, and Qi Wu. 2024b. T2VIndexer: A Generative Video Indexer for Efficient Text-Video Retrieval. In *ACM MM*.

Yuzhi Li, Haojun Xu, Feifan Cai, and Feng Tian. 2024c. Improving AI-assisted video editing: Optimized footage analysis through multi-task learning. *Neurocomputing*.

Zhiqiu Lin, Siyuan Cen, Daniel Jiang, Jay Karhade, Hewei Wang, Chancharik Mitra, Tiffany Ling, Yuhan Huang, Sifan Liu, Mingyu Chen, Rushikesh Zawar, Xue Bai, Yilun Du, Chuang Gan, and Deva Ramanan. 2025. Towards Understanding Camera Motions in Any Video. In *NeurIPS*.

Hongbo Liu, Jingwen He, Yi Jin, Dian Zheng, Yuhao Dong, Fan Zhang, Ziqi Huang, Yinan He, Yangguang Li, Weichao Chen, Yu Qiao, Wanli Ouyang, Shengjie Zhao, and Ziwei Liu. 2025. ShotBench: Expert-Level Cinematic Understanding in Vision-Language Models. In *NeurIPS*.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. 2021. Long-tail learning via logit adjustment. In *ICLR*.

Pamela A. Moss. 1994. Can There Be Validity without Reliability? *Educational Researcher*.

Isadora Campregher Paiva and Josephine Diecke. 2024. Revisiting Weimar Film Reviewers' Sentiments: Integrating Lexicon-Based Sentiment Analysis with Large Language Models. *Journal of Cultural Analytics*.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. 2021. Keeping Your Eye on the Ball: Trajectory Attention in Video Transformers. In *NeurIPS*.

Arianna Pera and Luca Maria Aiello. 2024. Narratives of Collective Action in YouTube's Discourse on Veganism. In *ICWSM*.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large Language Models Sensitivity to The Order of Options in Multiple-Choice Questions. In *NAACL Findings*.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. 2025. The Photographer's Eye: Teaching Multimodal Large Language Models to See, and Critique Like Photographers. In *CVPR*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *Preprint*, arXiv:2212.04356.

Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. 2020. A Unified Framework for Shot Type Classification Based on Subject Centric Lens. In *ECCV*.

Z. Rasheed and M. Shah. 2002. Movie genre classification by exploiting audio-visual features of previews. In *ICPR*.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. 2025. ARGUS: Hallucination and Omission Evaluation in Video-LLMs. In *ICCV*.

Mattia Savardi, András Bálint Kovács, Alberto Signoroni, and Sergio Benini. 2021. CineScale: A dataset of cinematic shot scale in movies. *Data in Brief*.

Mattia Savardi, András Bálint Kovács, Alberto Signoroni, and Sergio Benini. 2023. CineScale2: A dataset of cinematic camera features in movies. *Data in Brief*.

Gabriel S. Simões, Jônatas Wehrmann, Rodrigo C. Barros, and Duncan D. Ruiz. 2016. Movie genre classification with Convolutional Neural Networks. In *IJCNN*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning Large Multimodal Models with Factually Augmented RLHF. In *ACL Findings*.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. 2024. AI-Driven Review Systems: Evaluating LLMs in Scalable and Bias-Aware Academic Reviews. *Preprint*, arXiv:2408.10365.

Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. 2024a. Weakly Supervised Gaussian Contrastive Grounding with Large Multimodal Models for Video Question Answering. In *ACM MM*.

Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, and Raffay Hamid. 2023. Selective Structured State-Spaces for Long-Form Video Understanding. In *CVPR*.

Xinran Wang, Songyu Xu, Xiangxuan Shan, Yuxuan Zhang, Muxi Diao, Xueyan Duan, Yanhua Huang, Kongming Liang, and Zhanyu Ma. 2025. CineTechBench: A Benchmark for Cinematographic Technique Understanding and Generation. In *NeurIPS*.

Zhanyu Wang, Longyue Wang, Zhen Zhao, Minghao Wu, Chenyang Lyu, Huayang Li, Deng Cai, Luping Zhou, Shuming Shi, and Zhaopeng Tu. 2024b. GPT4Video: A Unified Multimodal Large Language Model for lnstruction-Followed Understanding and Safety-Aware Generation. In *ACM MM*.

Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards Long-Form Video Understanding. In *CVPR*.

Fanyi Xiao, Kaustav Kundu, Joseph Tighe, and Davide Modolo. 2022. Hierarchical Self-Supervised Representation Learning for Movie Understanding. In *CVPR*.

Junyu Xie, Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. AutoAD-Zero: A Training-Free Framework for Zero-Shot Audio Description. In *ACCV*.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024. MMAD:Multimodal Movie Audio Description. In *LREC-COLING*.

Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. 2023. Movie101: A New Movie Understanding Benchmark. In *ACL*.

Zihao Yue, Yepeng Zhang, Ziheng Wang, and Qin Jin. 2025. Movie101v2: Improved Movie Narration Benchmark. In *ACL*.

Emmanouil Zaranis, António Farinhas, Saul Santos, Beatriz Canaverde, Miguel Moura Ramos, Aditya K. Surikuchi, André Viveiros, Baohao Liao, Elena Bueno-Benito, Nithin Sivakumaran, Pavlo Vasylenko, Shoubin Yu, Sonal Sannigrahi, Wafaa Mohammed, Ben Peters, Danae Sánchez Villegas, Elias Stengel-Eskin, Giuseppe Attanasio, Jaehong Yoon, and 12 others. 2025. Movie Facts and Fibs (MF^2): A Benchmark for Long Movie Understanding. *Preprint*, arXiv:2506.06275.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025a. VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding. *Preprint*, arXiv:2501.13106.

Hongjie Zhang, Lu Dong, Yi Liu, Yifei Huang, Yali Wang, Limin Wang, and Yu Qiao. 2025b. LvBench: A Benchmark for Long-form Video Understanding with Versatile Multi-modal Question Answering. *IJCV*.

Qi Zhang, Zihao Yue, Anwen Hu, Ziheng Wang, and Qin Jin. 2022. MovieUN: A Dataset for Movie Understanding and Narrating. In *EMNLP Findings*.

Howard Zhou, Tucker Hermans, Asmita V. Karandikar, and James M. Rehg. 2010. Movie genre classification via scene categorization. In *ACM MM*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *Preprint*, arXiv:2504.10479.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*.

A Appendix

A.1 Prompts employed

For Pillars 1 and 2, we set up the tasks as a multiple-choice question-answer exercise with the following prompt:

Pillar 1-2 QA prompt

This video is a clip from (a movie | the movie <title>).
(This movie features <cast>).
(This movie is a <genre>).
(Who directed the movie? | What is the budget of this movie? | ...)
Pick one of the following options:
A) <option A>
B) <option B>
C) <option C>
Answer with the option's letter from the given choices directly.

For Pillar 3, we employed the following prompts to obtain scene descriptions for the quantitative and qualitative task respectively.

Pillar 3 Quantitative task prompt

You are a movie director. This is a scene from one of your movies, carefully describe what the core message of the scene is and which visual and stylistic elements contribute to that."

Pillar 3 Qualitative task prompt

You are a film scholar with a deep knowledge of film and cinematographic, and editing techniques. Provide a thorough and critical analysis covering different formal aspects of this movie clip. Be critical but fair. Conclude with three points of praise and three points of criticism. Only focus on the visuals. Refrain from discussing the sound and the title of the movie.
You should consider different aspects of what appears in the frame, including the setting, lighting, and costume. You should also consider the cinematography, like the different angles, shot compositions, and framing. Keep in mind aspects of the editing as well, such as the pace, rhythm, and transitions.

A.2 Pillar 1: Dataset descriptives

Here, we report a comparison in size and attributes of the existing LVU dataset and our proposed CMD-AD+.

Dataset	Attribute	# Videos	# Movies	# Classes
LVU	Director	107	12	8
	Genre	582	174	3
	Writer	168	18	7
	Year	141	108	9
CMD-AD+	Director	8687	1398	806
	Genre	8687	1398	439
	Country	8687	1398	21
	Budget	8117	1310	5

Table 5: Distribution of attributes across the splits of LVU and CMD-AD+. Compared to LVU, CMD-AD+ consists of more videos from a larger number of movies and it has more classes per attribute.

A.3 Pillar 1: Correlation between classification performance and popularity

	Video	Text	Video Text
Qwen2.5-VL	0.23	0.15	0.16
VideoLLaMA3	0.17	0.11	0.14
InternVL3	0.22	0.12	0.12
Gemma3n	0.19	0.12	0.16

Table 6: Correlation between F1-score for each director in CMD-AD+ and the director's average box-office.

	Video	Text	Video Text
Qwen2.5-VL	0.58	0.36	0.54
VideoLLaMA3	0.43	0.42	0.42
InternVL3	0.58	0.44	0.59
Gemma3n	0.56	0.57	0.53

Table 7: Correlation between F1-score for each genre class in CMD-AD+ and the genre's average box-office.

We conducted a preliminary analysis to investigate what could drive the difference in performance across genres and directors. We consider box office revenue as a proxy for fame and frequency in web-scale training data. We break down the results for each genre and director, and compute their correlation with their respective average box office revenue, as shown in Tables 6 and 7.

We observe a moderate positive correlation between revenue and genre F1-score, especially for the Video only and Video+Text setting across all models. This indicates that MLLM performance on this task is higher for genres with higher revenues and lower for less remunerating ones. The weaker correlation for the text-only setting indicates that less training data is needed compared to the video settings. On the other hand, the correlation between director F1 and director's box office revenue is weaker and more stable across the different settings.

A.4 Pillar 1 detailed tables

Table 8 and Table 9 show detailed results for the LVU and CMD-AD+ datasets, with performance per individual attribute with increasing amounts of context added in the textual modality. Across all models and attributes, we see that increasingly supplying more contextual information improves performance, showing that MLLMs need the knowledge supplied by the textual component to perform well.

A.5 Pillar 1 & 2: Visual encoder analysis

The MLLMs considered share similar architectures for their visual encoders (i.e., ViTs) with the exception of Gemma3n, which uses a CNN-based visual encoder. We find that the performance differences do not align with the architecture choice. For instance, in Table 1, Gemma3n performs better than InternVL3, a ViT-based MLLM with the same parameter count (300M). On the other hand, in Table 2 Gemma3n performs better than VideoLLaMa3 despite having fewer parameters dedicated to the vision encoder.

While for the first pillar, Table 1 and Table 2 suggest that a larger visual encoder is beneficial, as Qwen2.5-VL outperforms the other models, for the second pillar, Table 3 shows that Qwen2.5-VL does not outperform the other models, despite having the largest visual encoder. While the type of architecture and network size do not seem strong indicators for performance, we hypothesize that training data and training strategies may play a more important role. However, experimentally verifying this would be prohibitively expensive, given the scales of these models and the multiple stages of training involved.

	Vid.	Title	Cast	Dir.	Gen.	Writ.	Year	Avg.
VideoBERT (Sun et al., 2019)	✓			47.3	51.1	38.5	36.1	43.3
Obj_T4mer. (Wu and Krahenbuhl, 2021)	✓			47.7	52.7	36.3	37.8	43.6
Orthoformer (Patrick et al., 2021)	✓			55.1	55.8	47.0	43.4	50.3
VIS4mer (Islam and Bertasius, 2022)	✓			62.6	54.7	48.8	44.8	52.7
TranS4mer (Islam et al., 2023)	✓			63.9	55.9	46.9	45.5	53.1
S5 (Wang et al., 2023)	✓			67.3	65.4	51.3	48.0	58.0
Movies2Scene (Chen et al., 2023)	✓			70.9	55.9	53.7	57.8	59.6
VideoMamba (Li et al., 2024a)	✓			67.3	65.2	53.0	48.2	58.4
MA-LMM (He et al., 2024)	✓			74.6	61.1	70.4	51.9	64.5
HierarQ (Azad et al., 2025)	✓			78.4	67.9	71.9	61.9	70.0
Qwen2.5-VL (Bai et al., 2025)		✓		88.8	73.5	52.4	41.8	64.1
		✓	✓	86.9	78.4	54.8	51.8	68.0
	✓			85.0	73.5	59.5	53.9	68.0
	✓	✓		89.7	80.9	64.3	58.2	73.3
	✓	✓	✓	91.6	82.3	63.7	58.9	74.1
VideoLLaMA3 (Zhang et al., 2025a)		✓		57.9	66.5	58.9	41.1	56.1
		✓	✓	59.8	69.9	61.9	55.3	61.7
	✓			79.4	65.1	50.6	33.3	57.1
	✓	✓		85.0	70.4	59.5	51.1	66.5
	✓	✓	✓	88.8	73.0	63.7	62.4	72.0
InternVL3 (Zhu et al., 2025)		✓		85.0	80.9	66.1	52.5	71.1
		✓	✓	96.3	83.2	67.9	61.0	77.1
	✓			40.2	57.2	44.6	44.7	46.7
	✓	✓		82.2	77.8	62.5	60.3	70.7
	✓	✓	✓	94.4	79.6	65.5	64.5	76.0
Gemma3n (Google, 2025)		✓		79.4	74.1	57.7	36.9	62.0
		✓	✓	68.2	76.6	60.7	36.2	60.4
	✓			68.5	64.4	25.6	37.6	49.0
	✓	✓		84.1	74.9	57.1	38.3	63.6
	✓	✓	✓	82.6	75.8	66.1	38.3	65.7

Table 8: Movie attribute classification accuracy on the LVU dataset. The top half of the table reports on prior works which were evaluated only on video data, the lower half compares two general-purpose MLLMs in text-only, video-only, and multimodal settings.

	Video	Title	Cast	Other	Budget	Country	Director	Genre	Avg.
Qwen 2.5-VL		✓			33.9	22.3	34.8	61.2	38.0
		✓	✓		39.6	25.8	36.5	60.8	40.7
		✓	✓	✓	39.7	25.0	37.3	60.2	40.5
	✓				24.4 ± 2.3	15.9 ± 1.4	22.6 ± 0.4	45.9 ± 0.5	27.2
	✓	✓			37.9 ± 2.5	21.3 ± 2.2	35.8 ± 0.5	60.5 ± 0.4	38.9
	✓	✓	✓		39.9 ± 2.9	27.0 ± 0.8	37.7 ± 0.2	59.4 ± 0.4	41.0
	✓	✓	✓	✓	40.2 ± 2.9	28.2 ± 1.6	38.6 ± 0.4	59.9 ± 0.3	41.7
Video LLaMA3		✓			34.0	12.9	24.6	49.0	30.1
		✓	✓		37.3	22.6	26.6	49.7	34.0
		✓	✓	✓	38.3	27.6	26.9	51.4	36.1
	✓				28.4 ± 1.3	8.9 ± 0.6	10.7 ± 0.6	35.5 ± 0.4	20.9
	✓	✓			37.2 ± 2.9	13.4 ± 0.9	25.0 ± 0.6	50.1 ± 0.7	31.4
	✓	✓	✓		39.0 ± 1.9	21.5 ± 1.7	26.5 ± 0.8	47.5 ± 0.5	33.6
	✓	✓	✓	✓	38.1 ± 1.7	31.1 ± 2.4	27.6 ± 1.0	48.0 ± 1.0	36.2
InternVL3		✓			26.7	24.2	38.9	63.5	38.3
		✓	✓		27.9	24.0	43.3	65.2	40.1
		✓	✓	✓	27.5	24.6	43.2	66.2	40.4
	✓				27.6 ± 1.1	8.8 ± 0.9	19.9 ± 0.4	43.0 ± 0.4	24.8
	✓	✓			29.9 ± 1.3	25.8 ± 1.3	38.9 ± 1.4	60.9 ± 0.6	38.9
	✓	✓	✓		31.2 ± 1.4	27.6 ± 0.8	42.2 ± 0.9	62.0 ± 0.8	40.8
	✓	✓	✓	✓	34.7 ± 1.5	31.3 ± 3.1	44.2 ± 0.8	62.4 ± 0.9	43.1
Gemma3n		✓			37.9	21.5	39.7	61.0	40.0
		✓	✓		33.2	33.0	42.1	61.3	42.4
		✓	✓	✓	33.0	33.9	40.9	61.3	42.3
	✓				22.0 ± 0.7	12.0 ± 0.8	16.3 ± 0.4	46.4 ± 0.4	24.2
	✓	✓			27.1 ± 0.9	21.2 ± 2.3	32.6 ± 1.2	62.3 ± 0.2	35.8
	✓	✓	✓		27.0 ± 0.7	29.2 ± 5.1	32.1 ± 0.9	62.6 ± 0.4	37.7
	✓	✓	✓	✓	26.8 ± 0.5	34.7 ± 4.8	32.5 ± 1.1	63.0 ± 0.4	39.2

Table 9: Movie attribute classification F1-score on the CMD-AD+ dataset. CMD-AD+ contains multiple video clips per movie. As such, we report the average (and standard deviation) of randomly sampling a single video clip per movie 5 times. As the text-only settings do not consider the video clips, these are based on a single run.

	Vid.	Title	Cast	Director		Genre		Writer		Year	
				S	L	S	L	S	L	S	L
Qwen2.5-VL		✓		58.9	88.8	70.8	73.5	32.1	52.4	42.6	41.8
		✓	✓	59.8	86.9	66.8	78.4	28.0	54.8	52.5	51.8
	✓			86.0	85.0	70.3	73.5	58.3	59.5	48.2	53.9
	✓	✓		87.9	89.7	77.3	80.9	67.9	64.3	56.0	58.2
	✓	✓	✓	86.0	91.6	77.7	82.3	42.3	63.7	59.6	58.9
VideoLLaMA3		✓		39.3	57.9	60.1	66.5	33.3	58.9	22.7	41.1
		✓	✓	41.1	59.8	64.4	69.9	29.2	61.9	35.5	55.3
	✓			79.4	79.4	61.3	65.1	22.0	50.6	22.0	33.3
	✓	✓		85.0	85.0	69.2	70.4	40.5	59.5	32.6	51.1
	✓	✓	✓	85.0	88.8	68.2	73.0	32.7	63.7	37.6	62.4

Table 10: Model size comparison on LVU. We compare the smaller (S) versions of Qwen2.5-VL and VideoLLaMA3, 3B and 2B, respectively, against their larger (L) 7B variant.

	Angle		Motion		Size		Type	
	S	L	S	L	S	L	S	L
Qwen2.5-VL	40.5	45.8	29.1	28.8	52.7	55.7	43.8	53.6
VideoLLaMA3	46.9	48.6	29.1	30.5	46.5	50.2	49.2	58.6

Table 11: Model size comparison on AVE. We compare the smaller (S) versions of Qwen2.5-VL and VideoLLaMA3, 3B and 2B, respectively, against their larger (L) 7B variant.

A.6 Pillar 1 & 2: Model size comparison

To understand the influence of MLLM parameter size, we additionally compare the performance of smaller variants (3B and 2B for Qwen2.5-VL and VideoLLaMA3, respectively) to the 7B variants on LVU in Table 10. Across both model architectures, we predominantly notice a difference in performance between the smaller variants and the 7B variants in the text-only settings, where the reduction in model size leads to a drop in performance. Apart from writer classification and year classification for VideoLLaMA3, it appears that the reduction in model size is much less consequential for the settings involving video. Overall, we observe a positive effect from the larger model size across both architectures, but, from a multimodal perspective, it appears the visual encoder is a limiting factor for scaling the model size.

Similarly, as for LVU, Table 11 shows a comparison between MLLM versions of different sizes on AVE. We observe a drop in performance between 2 and 10 percentage points due to the reduction in model size across three out of four tasks. For motion classification, there is no benefit from the increase in model size. Moreover, the performance on this task is also the worst across all four tasks. We believe this may be caused by the models utilizing multiple frames rather than video, which may obfuscate the motion, and we expect fewer instances of discussion of motion in training data consisting of text and/or video.

A.7 Pillar 2: Further analysis on shot size

Among the cinematographic attributes of Pillar 2, prior research has predominantly focused on shot size, and it is frequently brought up in analyses of cinematography. To dive deeper into the results, we show the confusion matrix for shot size in Figure 5 for both VideoLLaMA3 and Qwen2.5-VL. Across models, we can observe that most confusions are between adjacent shot sizes, i.e., extreme close-ups are most often confused with close-ups, and extreme wides with wides. Interestingly, in almost all cases, extreme-wide shots are predicted as wide shots by VideoLLaMA3. However, possibly due to a lack of diversity in MLLMs' pretraining data, medium shots are most often predicted as close-ups by the models at hand. As the difference between adjacent shot sizes can sometimes depend on the definition used, it is promising that the MLLMs appear to understand the general difference between shot sizes.

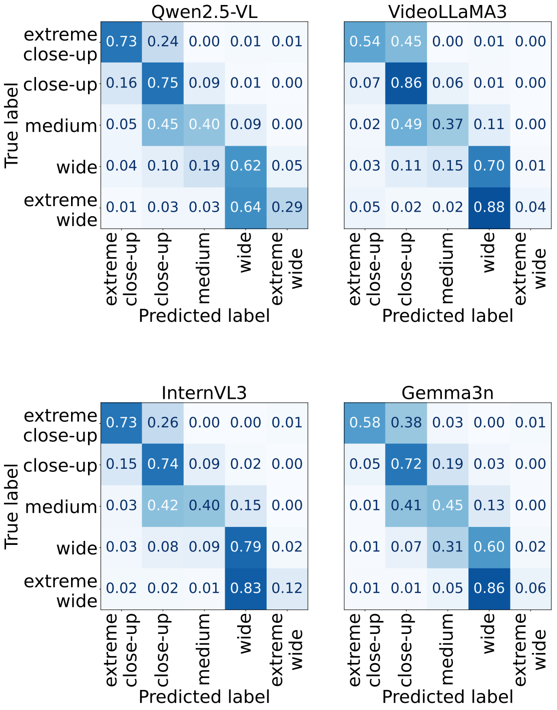


Figure 5: Confusion matrix for shot size for the MLLMs. Close-ups and their extreme counterparts are often mixed up. This is also the case for wide and extreme wide shots.

A.8 Pillar 3: Error rate breakdown

	Dynamic Action	Summ.	Visual Descr.
Qwen2.5-VL	83.5	54.0	92.8
VideoLLaMA3	85.9	76.7	95.3
InternVL3	85.8	60.5	90.5
Gemma3n	89.3	82.2	94.8

Table 12: Hallucination Error rates

To obtain a more fine grained look at the failure modes, we considered the categories used by the Argus framework to group statements. These are either summary statements, statements describing a static visual detail of the video, or statements describing dynamic actions, such as events or attribute/relationship changes. Each statement is then judged as entailed, contradictory, or undetermined, based on whether the statement is supported by the source material.
We show the detailed breakdown in Tables 12 and 13. Both in terms of hallucination and omission, the summary statements have the lowest error rates.

	Dynamic Action	Summ.	Visual Descr.
Qwen2.5-VL	92.0	85.3	89.4
VideoLLaMA3	95.9	95.0	96.2
InternVL3	95.2	94.9	94.2
Gemma3n	94.7	90.7	92.2

Table 13: Omission Error rates

On the other hand, for hallucination, dynamic actions have slightly lower error rates than visual descriptions, while they have similar error rates for omission. Overall, this hints that MLLMs are able to make more general statements that are applicable, but that they severely struggle with the video-specific understanding required for the task. Improved mechanisms to ensure that statements are visually-grounded and supported by the provided context are necessary for this task.