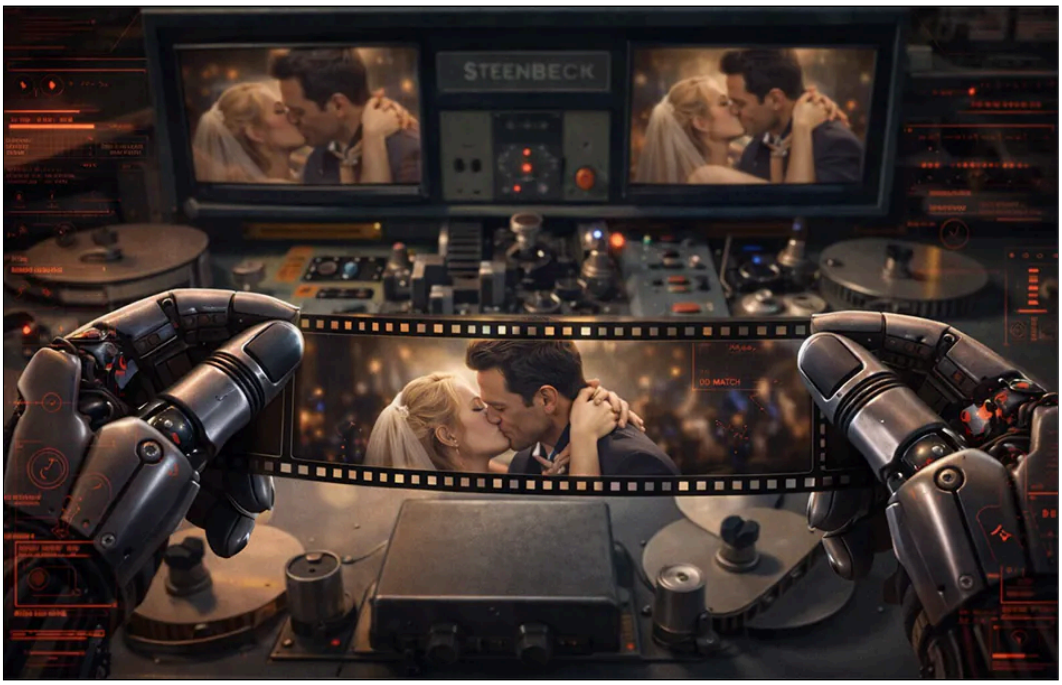
In Search of an AI That Can Follow an Entire Movie

Originally published at https://www.unite.ai/in-search-of-an-ai-that-can-follow-an-entire-movie/



AI models still lose track of who is who and what's happening in a movie. A new system orchestrates face recognition and staged summarization, keeping characters straight, and plots coherent across full-length films.

Getting artificial intelligence to watch and understand Hollywood-style movies may seem like a niche or marginal pursuit; but a system that can watch a full-length film from beginning to end, track the progress of all the characters, and stay on top of the plot, has not only made possible a number of direct applications that could benefit from such capabilities, but also a fair few peripheral or unrelated challenges, across different domains.

The low-hanging fruit for movie-watching AI models is recommender systems, in streaming platforms such as Netflix, Amazon Prime, and HBO Max. A granular understanding of plot developments and character-actions allows for a closer match to the (often-specious) predilections and enthusiasms of viewers.

Further, a deeper comprehension of a film enables keyword generation and more accurate categorization, rather than perpetuating oft-copied movie descriptions that may have been written decades ago. Such insights could also surface the presence of 'adult' themes in a movie which may not be obvious from dialogue, or from the visuals.

Additionally, older movies in a catalogue may hold outdated ratings, as well as overviews; for instance, language and idioms that were normalized in a 1950s movie might warrant far more attention now. But without an overall understanding of context, gleaned from genuinely following a long movie narrative, such incidences could be over or under-emphasized.

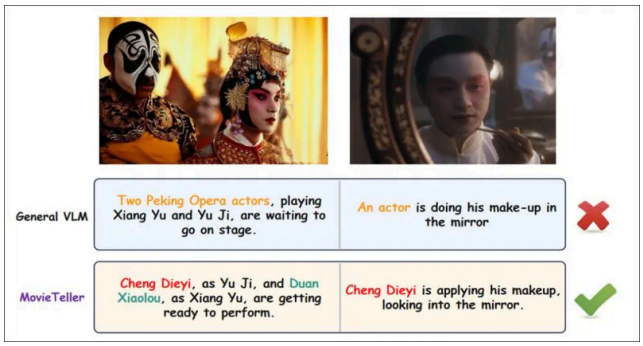
More broadly, improved movie analysis approaches could contribute a great deal to the much wider-ranging problem of event recognition, which is vitally needed for innovations in security monitoring, automated sports commentaries, and summaries of all kinds, across a huge range of media.

Therefore, 'AI-based movie-watching' is a surprisingly well-subscribed genre in Computer Vision literature.

Seeing the Big Picture

The latest entrant is titled *MovieTeller* – an academic/industry collaboration from China that makes new headway by dividing up the various sub-tasks in the challenge across various AI applications that suit these challenges, instead of – as is often the case – attempting to train discrete and encapsulated models that can perform all the necessary tasks from a single latent space.

The authors observe that prior Vision-Language Models (VLMs) faced with the same task have not been able to progress much beyond single-frame analysis; and that their lack of context makes it hard for such models to persistently identify characters – perhaps the most essential characteristic of such a system:



The new system, MovieTeller, is able to persistently identify people in scenes, thanks to the use of a dedicated facial recognition system; but it's the more overarching dedication to context that allows the framework to keep on top of plot developments. Source

The authors state:

> *'General-purpose VLMs often struggle to recognize and consistently track specific characters throughout a lengthy narrative. They may describe a key protagonist as "a man" in one scene and "a person" in another, failing to bind the visual representation to a consistent identity.'*

The authors note that because Transformers' self-attention mechanism makes use of quadratic complexity, processing every frame of a full-length movie at once becomes too computationally expensive. As a result, approaches that rely on uniform frame-sampling or simple concatenation tend to break up the flow of the story, producing fragmented summaries instead of a coherent narrative.

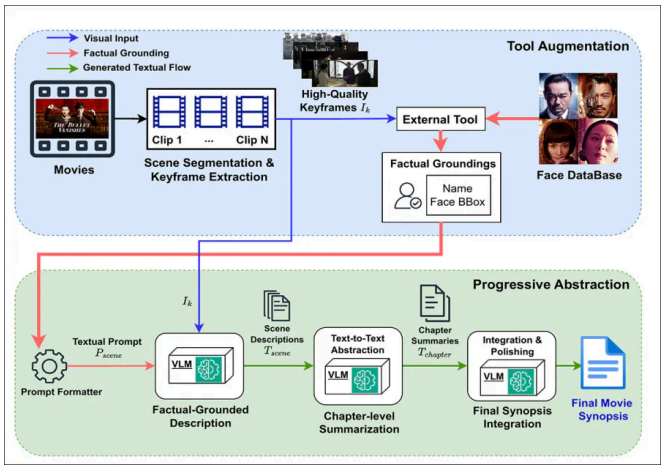
Instead, the new system comprises an orchestrated training-free pipeline, with dedicated tools to address face recognition and persistence of memory (as characters leave and re-enter the narrative of a movie).

MovieTeller was tested against prior approaches using 60 full-length movies, equivalent to 10,000 minutes of footage. In quantitative ablation tests and human studies, the authors report, their approach was able to improve notably on the default environments and assumptions used by prior systems.

The new paper is titled *MovieTeller: Tool-augmented Movie Synopsis with ID Consistent Progressive Abstraction*, and comes from five authors across Zhejiang University at Hangzhou, the state-led China Media Group, and Watch AI Group* (the latter two based in Beijing).

Method

The MovieTeller schema comprises three stages: scene segmentation and keyframe extraction, which are handled through the PySceneDetect project; Factual-Grounded Scene Description Generation via the customization of the Qwen2.5-VL-7B-Instruct VLM; and *progressive abstraction*, which condenses detailed scene descriptions into chapter summaries, and then into a final coherent synopsis – and this is also performed by the Qwen2.5 model:



Overview of the MovieTeller framework: a full-length film is first segmented into scenes and distilled into high-quality keyframes; then, an external face-recognition tool injects factual groundings, linking character names to bounding boxes, which guide a Vision-Language Model in producing ID-consistent scene descriptions. These descriptions are then progressively abstracted into chapter summaries and integrated into a coherent final movie synopsis.

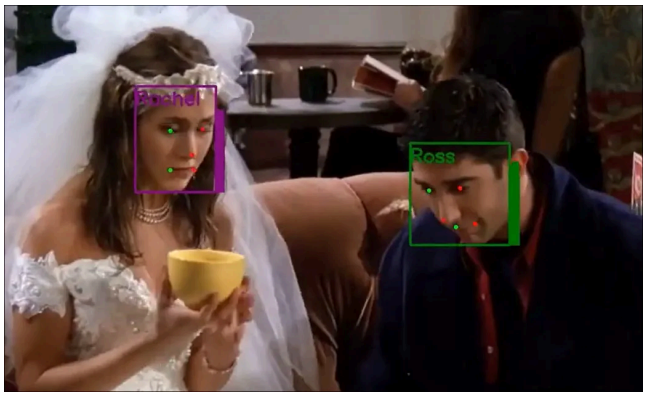
The initial stage uses PySceneDetect to break the film into discrete scenes, based on clear visual changes, with each scene represented by a single keyframe.

However, not every frame makes a good summary image, since transitional moments, fade-outs, and dark frames can confuse later analysis. Therefore a simple quality-check performs a filter run on candidate frames, by measuring brightness and visual variation, ensuring that only information-rich images are selected for description.

Placing the Face

A face database was built from publicly-available cast information†, storing each main character's name alongside a numerical face embedding. When a face appears in a keyframe, its embedding is matched against the database, and the closest result accepted if it clears a confidence threshold. This creates 'factual groundings', linking names to specific bounding boxes.

For these purposes, InsightFace is used, leveraging an ArcFace loss-based recognition head:



Two familiar faces well-remembered by the Additive Angular Margin Loss (ArcFace) initiative, used in a very similar way for the MovieTeller project. Source

The annotated keyframes are then passed to the Qwen model with a prompt that lists detected characters and their positions.:

Since Vision-Language Models can't absorb an entire feature-length film at one pass, MovieTeller initially breaks down the material into scene descriptions. These are grouped into consecutive, chapter-like blocks, which are subsequently passed to Qwen2.5, which summarizes each chapter, compressing plot developments, character motivations and turning points, while retaining the previously-verified character names.

Those compressed chapter summaries are then concatenated and returned to the model with a new prompt that solicits a complete synopsis:

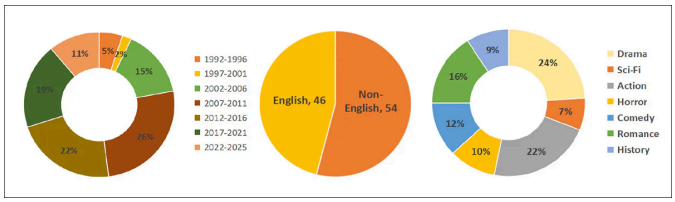
```
You are a professional movie analyst. Please generate a
    summary for the movie scene represented by the
    following keyframe.
To assist you, I have identified the following characters
    and their locations in the image:
- The actor '{Character_Name_1}' is located within the
    bounding box {BBox_1}.
- The actor '{Character_Name_2}' is located within the
    bounding box {BBox_2}.
...

Please generate the scene summary by combining the
    character identities with the visual content:
```

Similar to the prompt that asks for an entire synopsis, this sample is used to generate scene descriptions*, explicitly injecting verified character names and bounding boxes to constrain the Vision-Language Model, and enforce ID-consistent narration.*

Assuming the process has succeeded, the final output should cohesively reflect the film's narrative arc. This is a particularly tough task in machine learning, since the variety of possible plot summaries, and the style in which they might be presented, together with the necessary length of these data points, makes it pretty much impossible to adopt the usual ground truth-based approaches.

Data and Tests

To test the system, the authors curated a bespoke (and source-unattributed) dataset of 100 full-length movies, equivalent to around 166 hours of runtime. The movies included *Iron Man 3*, *Farewell My Concubine*, *Eat Drink Man Woman*, and *The Chronicles of Narnia*. The researchers required that all included movies rate above 5.0 on the IMDB:



Dataset composition across 100 films, showing balanced temporal coverage from 1992 to 2025, a slight majority of non-English titles, and a broad genre spread led by Drama and Action, with representation across Sci-Fi, Horror, Comedy, Romance, and History.

The wide range of genres addressed (see graph above) was designed to prevent bias towards any one genre.

The face dataset for each film comprised of two pictures of principal actors – one from a film still, and one from a related publicity photograph.

Implemented in Python, the tests were run across four NVIDIA A40 GPUs, each with 48GB of VRAM, and with the aforementioned Qwen2.5 variant as the central VLM. Ablation studies†† were also conducted with alternate state-of-the-art models InternVL3-8B and WeThink-Qwen2.5VL-7B.

The new framework was tested against two ablation†† variants: a *No-Hint* baseline, in which the Vision-Language Model generated scene descriptions from the keyframe alone, without any textual cues about character identities; and a *Name-Only Hint* setting, where the model was given the detected character names, but not their bounding boxes, allowing the authors to isolate the specific contribution of spatial grounding to identity consistency and narrative coherence

In regard to metrics, considering the aforementioned difficulty of applying ground-truth methods to long plot summaries, standard *n*-gram overlap metrics such as ROUGE and BLEU were eschewed in favor of BERTScore with F1 score, to measure semantic similarity against a reference synopsis drawn from 'a public encyclopedia'.

Also, Gemini 2.5 Flash was used to score each synopsis for factual faithfulness; ID consistency and completeness; narrative coherence and flow; and *conciseness*, with scores averaged across dimensions.

Finally, a human evaluation of 50 randomly sampled summaries was conducted using pairwise comparison, providing a practical check on the automated assessments.

Below we see BERTScore (F1) results for the three backbone models: Qwen2.5-VL, InternVL3, and WeThink. Each is tested in three configurations: *No-Hint*, *Name-Only*, and the full MovieTeller system:

Method	Qwen2.5-VL	InternVL3	WeThink
No-Hint	0.612	0.616	0.618
Name-Only	0.616	0.617	0.624
MovieTeller (Ours)	**0.638**	**0.631**	**0.639**

BERTScore (F1) comparison across three base Vision-Language Models and three experimental settings, showing consistent gains from adding character names and further improvements when spatial grounding is included, with MovieTeller achieving the highest scores in all cases.

The authors note that the pattern is consistent across all three backbone models: using only the raw keyframe yields the weakest performance; adding character names produces a modest improvement; and combining names with bounding boxes delivers the strongest results. Although the gains are incremental rather than dramatic, the fully-grounded configuration achieves the highest semantic alignment with the reference synopsis, in every setting.

In regard to LLM-based evaluation of narrative quality: as we see in the results below, the *No-Hint* baseline struggles most with identity consistency, which pulls down its overall score; but supplying names alone produces a noticeable lift, particularly on identity-related dimensions. However, the full MovieTeller configuration again ranks highest across all three backbone models:

Method	Faith.	ID Cons.	Coherence	Concise.	Final
Base Model: Qwen2.5-VL-7B-Instruct					
No-Hint	2.49	1.80	2.24	2.17	2.18
Name-Only	3.15	3.55	2.35	2.41	2.87
MovieTeller (Ours)	**3.21**	**3.65**	**2.52**	**2.45**	**2.96**
Base Model: InternVL3-8B					
No-Hint	2.51	1.75	2.20	2.21	2.17
Name-Only	3.22	3.67	2.26	2.38	2.88
MovieTeller (Ours)	**3.34**	**3.80**	**2.50**	**2.44**	**3.02**
Base Model: WeThink-Qwen2.5VL-7B					
No-Hint	2.45	1.82	2.26	2.25	2.20
Name-Only	3.19	3.65	2.38	2.42	2.91
MovieTeller (Ours)	**3.30**	**3.76**	**2.55**	**2.51**	**3.03**

LLM-as-a-Judge evaluation (1–5 scale) across three base models, showing that adding character names improves identity consistency and overall quality, while the full MovieTeller framework achieves the highest scores across factual faithfulness, coherence, conciseness, and final rating.

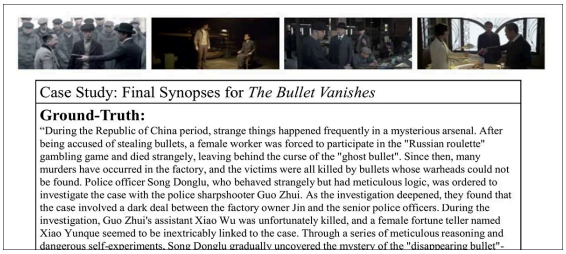
The strongest gains appear in ID consistency, and in the final averaged score, suggesting that spatial grounding helps the model stay clear about who is doing what as the plot unfolds.

In the human evaluation of 50 randomly sampled summaries, participants were shown three summaries at a time, and asked to select the best:

Method	Qwen2.5-VL	InternVL3	WeThink
No-Hint	6%	4%	6%
Name-Only	34%	42%	32%
MovieTeller (Ours)	**60%**	**54%**	**62%**

Human preference rates in a three-way forced-choice evaluation, showing that the fully grounded MovieTeller summaries are selected most often across all three base models, significantly outperforming both the No-Hint and Name-Only variants.

Finally, a qualitative test was run, on the film *The Bullet Vanishes* (2012):



Case Study: Final Synopses for *The Bullet Vanishes*

Ground-Truth:

"During the Republic of China period, strange things happened frequently in a mysterious arsenal. After being accused of stealing bullets, a female worker was forced to participate in the "Russian roulette" gambling game and died strangely, leaving behind the curse of the "ghost bullet". Since then, many murders have occurred in the factory, and the victims were all killed by bullets whose warheads could not be found. Police officer Song Donglu, who behaved strangely but had meticulous logic, was ordered to investigate the case with the police sharpshooter Guo Zhui. As the investigation deepened, they found that the case involved a dark deal between the factory owner Jin and the senior police officers. During the investigation, Guo Zhui's assistant Xiao Wu was unfortunately killed, and a female fortune teller named Xiao Yunque seemed to be inextricably linked to the case. Through a series of meticulous reasoning and dangerous self-experiments, Song Donglu gradually uncovered the mystery of the "disappearing bullet"-

We cannot reproduce the entirety of this figure from the original paper, since it is very tall and text-dense. Please refer to the source paper instead.

Here, the *No-Hint* baseline produces a vague synopsis that refers to characters in generic terms, and blurs their roles, making the chain of events harder to follow. Supplying names alone improves surface-level recall, but the narrative still drifts, with character relationships and motivations described in a rather 'flattened' way.

Conversely, the fully grounded MovieTeller version keeps identities stable throughout the synopsis and ties actions to the correct characters, allowing the investigation plot to unfold with clearer causal structure. Specific tensions and role dynamics are preserved rather than abstracted away, resulting in a summary that reads less like a disconnected outline and more like a coherent retelling of the film's central arc:

Name-Only Hint (Ablation):

"In a city full of suspense and danger, detective Guo Zhui, dressed in black and holding a silver gun, calmly investigates a series of domestic violence cases, and gradually uncovers the complex conspiracy behind them with the help of his partner Song Donglu. Chief Jin carefully directs behind the scenes, hinting at the intervention of senior management; assistant Xiao Wu bravely supports, bringing warmth and hope; the mysterious woman Xiao Yunque provides key clues, and her identity is a mystery but crucial. As the fate of the five people interweaves, the truth gradually emerges. In the end, the three parties join forces to fight the mastermind behind the scenes, launching a life-and-death duel, justice is upheld, and everyone's fate is changed."

MovieTeller (Ours):

"Against a backdrop of suspense and crisis, detective Guo Zhui works closely with key character Song Donglu to investigate a series of intricate domestic violence cases. Guo Zhui is calm and focused, wielding a silver gun through dangerous situations, while Song Donglu performs secret missions in the dark, and the two work together to find the truth in dim ancient buildings. Their investigation is directed by the authoritative and cautious Director Jin behind the scenes, but "Director Jin's intention" suggests that there may be a deeper conspiracy behind the scenes, making the whole process full of obstacles and dangers. With the help of the brave assistant Xiao Wu and the mysterious woman Xiao Yunque, the case becomes more and more confusing, with clues and mysteries intertwined. As their respective destinies are closely linked, they face tense confrontations and deadly dangers together. In the end, all the clues converge, Guo Zhui, Song Donglu and Director Jin join forces to fight the evil forces to the death, successfully revealing the shocking truth that has been hidden for many years, and ushering in the final victory of justice."

Part of the final comparison, which we cannot reproduce in full here, showing an ablated and a full MovieTeller summary. Please refer to the source paper instead.

Conclusion

Though most new projects of this kind end up in the Computer Vision literature, AI-generated film summarization encompasses many other disciplines and domains in machine learning research – and it is hard to tell which of these will unwittingly contribute the missing piece of the puzzle; though MovieTeller takes a step in the right direction, by splitting up the tasks across apposite modules instead of hoping to solve it all discretely in the latent space, it retains the 'cobbled-together' feel that tends to precede a later, more elegant solution.

* *I cannot identify this institution, even after some searching.*

† *One would presume something like the IMDB or OMDB, but the source is not specified.*

†† *Please refer to the source paper for comprehensive ablation, as we only cover full ablation in exceptional cases. I will note that the un-treated ablation studies referred to here do not undermine the paper's general findings.*

First published Friday, February 27, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai