OmniScript: Towards Audio-Visual Script Generation for Long-Form Cinematic Video

Junfu Pu*, Yuxin Chen*, Teng Wang*, and Ying Shan

ARC Lab, Tencent, China
{jevinpu,uasonchen}@tencent.com

Abstract. Current multimodal large language models (MLLMs) have demonstrated remarkable capabilities in short-form video understanding, yet translating long-form cinematic videos into detailed, temporally grounded scripts remains a significant challenge. This paper introduces the novel video-to-script (V2S) task, aiming to generate hierarchical, scene-by-scene scripts encompassing character actions, dialogues, expressions, and audio cues. To facilitate this, we construct a first-of-its-kind human-annotated benchmark and propose a temporally-aware hierarchical evaluation framework. Furthermore, we present *OmniScript*, an 8B-parameter omni-modal (audio-visual) language model tailored for long-form narrative comprehension. OmniScript is trained via a progressive pipeline that leverages chain-of-thought supervised fine-tuning for plot and character reasoning, followed by reinforcement learning using temporally segmented rewards. Extensive experiments demonstrate that despite its parameter efficiency, OmniScript significantly outperforms larger open-source models and achieves performance comparable to state-of-the-art proprietary models, including Gemini 3-Pro, in both temporal localization and multi-field semantic accuracy. The code is available at `https://github.com/TencentARC/OmniScript`.

Keywords: Video Script Transcription · Video Understanding · MLLM

1 Introduction

The analysis of cinematic content is a complex cognitive task, requiring the interpretation of a rich tapestry of visual and linguistic elements, including character relationships, narrative progression, and dialogue nuances. A deep understanding of these components is not only a cornerstone of computational media analysis but also holds significant practical value for the film and television industry, with potential applications in automated logging, content retrieval, and assisting human creativity. While recent advancements in multimodal large language models (MLLMs) [7, 15, 28, 34] have shown remarkable promise in video understanding, they predominantly focus on short-form video clips and tasks such as captioning or question-answering [8, 22, 30, 33]. The more ambitious and practical goal

* Equal contribution.

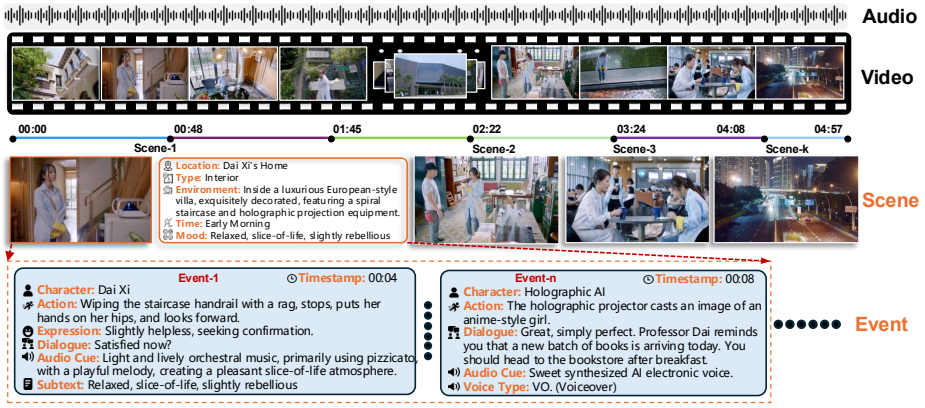

Fig. 1: Overview of our Video-to-Script (V2S) framework. Given a long-form cinematic video, our pipeline performs temporally grounded scene-event parsing and generates a structured script with multimodal fields (dialogue, action, expression, and audio cues).

of translating a full-length movie or episode into a detailed, structured script remains a largely unexplored frontier.

In this paper, we introduce the first Video-to-Script (shown in Fig. 1) multimodal large language model, a novel omni system designed to ingest long-form videos, ranging from several to tens of minutes, and generate a comprehensive, scene-by-scene script. This generated output includes not only the visual setting, such as the location, time of day, and atmosphere of each scene, but also a detailed breakdown of the in-scene narrative: character actions, dialogues (including voiceovers), and emotions. This task, which we term "holistic script generation," pushes the boundaries of current video MLLMs by demanding fine-grained, long-range temporal understanding and coherent narrative synthesis.

However, the development of such a model is fraught with significant challenges. First, there is a critical lack of suitable training data. Annotating long-form videos with the necessary granularity, parsing complex multi-scene structures and intricate character interactions, is an exceptionally labor-intensive and time-consuming endeavor. This data scarcity poses a fundamental obstacle to training models capable of deep narrative comprehension. Second, evaluating the quality of generated scripts presents a unique metrological challenge. Unlike tasks with single ground-truth answers, the model's output is a long, open-ended, and richly structured narrative containing multiple elements with precise temporal stamps. Defining robust metrics that capture the accuracy, coherence, and completeness of such complex generations remains an open problem. Third, the autoregressive generation of such detailed descriptions is computationally prohibitive. Our preliminary analysis indicates that describing a mere two-minute clip requires approximately 4,000 tokens. As the video duration extends, the token count—and consequently the inference cost and time—can explode, creating a critical bottleneck for practical efficiency.

To overcome these hurdles, we propose a comprehensive and novel solution consisting of three key contributions. **First**, we introduce the first-of-its-kind movie and TV series understanding dataset. This dataset encompasses a diverse range of content, including both horizontal-format films and vertical-format short dramas, covering a wide array of themes and genres to ensure robust model training. **Second**, we establish a dedicated, human-annotated benchmark and a suite of evaluation metrics specifically designed for the video-to-script task. By rigorously assessing the performance of existing closed-source and open-source MLLMs, our benchmark not only quantifies the current state-of-the-art but also highlights the significant gap in research and capability that our work aims to fill. **Third**, we propose the first omni-modal (audio+visual) script generation model. We leverage extensive audio-visual pre-training, CoT-style SFT with plot and character relationship reasoning, and employ an open-ended reward for script quality. This reward is effectively utilized in a GRPO-based post-training phase to align the model with human preferences for coherent and accurate storytelling.

Extensive experiments demonstrate the superiority of our approach. Our model not only achieves performance comparable to state-of-the-art closed-source models, including Gemini 3-Pro, on long-form video script generation tasks. This work represents a significant step toward automating the intricate process of video understanding and narrative generation, opening new avenues for research and application in computational media analysis.

2 Related Work

2.1 Cinematic Video Understanding and Narration

Early cinematic understanding evolved from short-clip descriptions [19,20,23,26] to plot-centric reasoning [4, 13, 25] and structural annotations (e.g., AVA [10], MovieGraphs [27], MovieNet [11]). However, a persistent dichotomy remains: macro-plot analyses lack fine-grained temporal grounding, while structural annotations fail to yield readable, coherent narratives. Recent efforts like Movie101 [31, 32] generate role-aware narrations but still predominantly operate on isolated, dialog-free clips, failing to model full-length film intricacies. In contrast, our proposed Holistic Script Generation task requires transcribing full-length videos into temporally anchored, hierarchically structured scripts. Unlike existing datasets [20,23,31] that entangle multimodal cues into coarse paragraph summaries, our annotation format strictly decouples fine-grained atomic elements, such as individual actions, expressions, dialogues, and audio cues, providing a much more rigorous benchmark for long-term narrative comprehension.

2.2 Dense and Omni-Modal Video Captioning

Traditional dense video captioning focuses on localizing salient events but typically yields sparse, concise summaries that neglect rich multimodal nuances [9,12,17]. Conversely, recent advancements in audio-visual captioning (e.g., AVoCaDO [5], video-SALMONN-2 [24], and DiaDem [6]) achieve deep multimodal

integration but generate holistic descriptions devoid of explicit temporal grounding. Empowered by advanced MLLMs, bridging this gap has emerged as a new frontier. Notably, a concurrent work, TimeChat-Captioner [29], introduces an "Omni Dense Captioning" task to generate timestamped, scene-level descriptions. While TimeChat-Captioner pioneers macro-scene structuring, its character actions, intents, and dialogues remain entangled within coarse summaries. Our OmniScript framework transcends these limitations by enforcing a strictly decoupled, hierarchical structure (*Scene* $\rightarrow$ *Event* $\rightarrow$ *Field*) that explicitly isolates fine-grained atomic elements anchored to dense timestamps, effectively resolving the ambiguity present in prior captioning paradigms.

3 Video-to-Script Generation

In this section, we first define the cinematic video-to-script problem and output schema, then describe the memory-augmented automatic annotation pipeline used to construct high-quality training data.

3.1 Problem Definition

Our task focuses on cinematic video understanding, where the goal is to transcribe a long-form movie/TV video into a temporally grounded, hierarchically structured script. Given an input video V with duration T, the model predicts a structured output $\hat{Y} = \{\hat{M}, \hat{S}\}$, where $\hat{M}$ denotes video-level metadata, and $\hat{S}$ denotes scene-event scripts.

Input. The input is an untrimmed cinematic video V containing complex narrative transitions, recurring characters, and multimodal cues (visual, speech, sound effects, and background music).

Output Schema. The output follows a JSON-style schema with three levels:

- **Meta-level ($\hat{M}$):** global attributes, including title/description, total duration, and a character list.
- **Scene-level ($\hat{S}$):** a sequence of scenes $\{s_i\}_{i=1}^{N_s}$, where each scene contains a scene identifier, location, and coarse time attribute (e.g., day/night).
- **Event-level ($\hat{E}_i$):** each scene s_i contains ordered events $\{e_{i,j}\}_{j=1}^{N_i}$ with timestamp and character identity, while each event includes at least one content field from {dialogue, action, expression, audio cue (e.g. sound event or BGM)}.

Formally, for an event $e_{i,j}$ occurring at time $\tau_{i,j}$, we define

$$e_{i,j} = (\tau_{i,j}, c_{i,j}, d_{i,j}, a_{i,j}, x_{i,j}, u_{i,j}), \tag{1}$$

where $c_{i,j}$ is character (or `Environment`), and $d/a/x/u$ denote dialogue/action/expression/audio cue, respectively.

Objective. The overall objective is to learn a mapping $f_\theta : V \rightarrow \hat{Y}$ that jointly optimizes: (1) temporal localization accuracy (when events happen), (2) character-consistent semantic parsing (who does/says what), and (3) multimodal narrative faithfulness (how the event is expressed through language, behavior, expression, and sound).

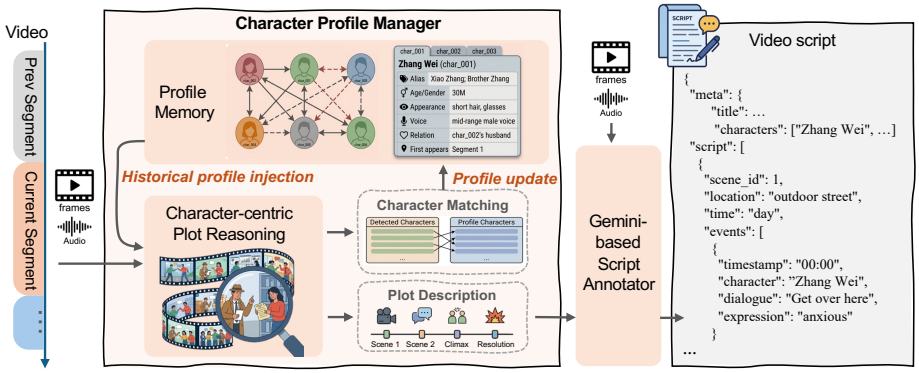

Fig. 2: Overview of the memory-augmented progressive annotation pipeline. The character profile manager injects historical profiles to guide plot reasoning and dynamically updates character memory. The generated plot description and raw audio-visuals are then fed into a Gemini-based annotator to produce a fine-grained video script.

3.2 Memory-Augmented Annotation Pipeline

Cinematic videos are characterized by intricate narratives, complex character dynamics, and long-range temporal dependencies. Traditional annotation pipelines, which independently process video segments and subsequently merge the textual outputs, often fail to maintain narrative coherence. Furthermore, character re-identification across scenes remains a critical bottleneck, as discriminative cues in movies are inherently multimodal (e.g., speech tone, gait, clothing) rather than purely facial. To overcome these limitations, we propose a novel, memory-augmented progressive annotation pipeline centered around a character profile manager (CPM). This pipeline efficiently processes raw videos into high-quality, chain-of-thought formatted script data. As illustrated in Fig. 2, our pipeline operates through three progressive stages:

Memory-Augmented CPM. Given a raw video, we first partition it into semantically cohesive segments (typically $<$5 minutes) using PySceneDetect based on scene boundaries. From a curated collection of over 10K raw cinematic videos (ranging from minutes to hours), we extract approximately 45K segments. For each segment, we employ a short-video expert model (i.e., Gemini-2.5-Pro) to perform character-centric plot reasoning. Crucially, this process is conditioned on the CPM, a memory module that stores cross-segment character information. By integrating current audio-visual inputs with historical profiles retrieved from the CPM, the model accurately resolves character identities and synthesizes a coherent plot description. Simultaneously, the CPM dynamically updates to ensure global consistency by continuously accumulating evolving multimodal attributes (e.g., costume changes) into existing character profiles. Furthermore, it employs a lazy naming strategy for entity resolution, where provisional char-

acter IDs are retroactively upgraded to permanent ones and duplicate records are merged upon detecting definitive naming events in the dialogue.

Fine-Grained Script Generation. With globally consistent plot descriptions established, we proceed to detailed script generation. The generated plot descriptions, which now encapsulate accurate character identities and narrative logic, are fed alongside the original audio-visual content into a powerful MLLM (i.e., Gemini). This step translates the high-level plot into a fine-grained, scene-by-scene script, capturing fine-grained audio-visual cues like character actions, dialogues, and emotional nuances.

Thinking Data Construction. To endow our target model with robust reasoning capabilities, we construct CoT trajectories driven by plot and character dynamics. We utilize a strong LLM (i.e., DeepSeek) to retroactively distill an "intermediate thinking" process from the generated scripts. This thinking phase explicitly articulates plot summaries and character relationship mappings. Consequently, we formulate a structured Video $\rightarrow$ Thinking $\rightarrow$ Script CoT dataset. This structured data not only provides high-quality supervision for script generation but also serves as the foundation for our model's reasoning-based training.

4 The V2S Benchmark: Metrics and Dataset

Evaluating highly structured, temporally grounded, and semantically rich video scripts poses a unique challenge. Traditional metrics (e.g., BLEU [16], ROUGE [14], or standard temporal action localization metrics) fail to capture the hierarchical dependencies and the open-vocabulary descriptions. To bridge this gap, we introduce a comprehensive Video-to-Script (V2S) benchmarking.

4.1 Hierarchical Evaluation Metrics

We devise a temporally-aware evaluation framework that rigorously assesses both structural correctness and semantic fidelity at the event and scene levels. Taking event-level evaluation as an example, our pipeline operates through four stages:
1. Text-Content-Guided Event Alignment. Relying solely on temporal Intersection-over-Union (tIoU) heavily penalizes semantically correct but temporally shifted predictions. To address this, we compute a composite semantic similarity score between candidate sequences of Ground Truth (GT) and predicted events. By leveraging dynamic programming (DP), we formulate a Weighted Interval Scheduling problem to efficiently resolve the optimal, order-preserving assignment under temporal proximity and non-overlap constraints.
2. LLM-Assisted Semantic Character Resolution. To disambiguate open-vocabulary character identities (e.g., matching a predicted "man in white" to the GT "John"), we employ an LLM to parse extracted entities and determine their specific roles or pluralities. We then construct an optimal bipartite mapping

graph G enforcing strict semantic constraints, preventing conflicting identity assignments.

3. Multi-dimensional Field Evaluation. We assess content quality ("what happens") by computing field content similarity across the aligned groups. Semantic fields (e.g., action, mood) use an LLM to measure semantic similarity S_{group} between predicted and ground truth field content; lexical fields (e.g., dialogue) use normalized Levenshtein distance; and categorical fields (e.g., character) use exact string matching via graph G. Given the total number of specific field in prediction (N_{pred}) and ground truth (N_gt), the global Precision ($P_{field} = \frac{\sum S_{group}}{N_{pred}}$) and Recall ($R_{field} = \frac{\sum S_{group}}{N_{gt}}$) are aggregated to calculate final harmonic mean as $F1_{field} = \frac{2 \cdot P_{field} \cdot R_{field}}{P_{field} + R_{field}}$.

4. Temporal Boundary Evaluation (tIoU Hit Rate). To rigorously assess localization quality of the predicted timestamps ("when it happens"), a predicted event group is considered a True Positive at a strictness threshold t (where $t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$) if its overlap with the GT group satisfies $tIoU \geq t$. We compute temporal precision ($P_{time}@t$) and recall ($R_{time}@t$) across all groups, resulting in the final metric $tIoU@t = \frac{2 \cdot P_{time}@t \cdot R_{time}@t}{P_{time}@t + R_{time}@t}$. More details about evaluation metric can be found in supplementary materials.

4.2 Construction and Quality Assurance

To systematically evaluate models, we construct a meticulously curated benchmark of 10 full-length cinematic works (19.9 hours total) covering diverse genres (anime, action, suspense, drama). Unlike traditional flat video captioning datasets, our benchmark introduces a dense, hierarchical structure. It decomposes into 1.4k distinct scenes and over 16.8k structural events, averaging an exceptionally high density of 14.1 events per minute. Crucially, it incorporates nuanced modalities rarely present in existing benchmarks, such as facial expressions, audio cues, and subtexts. To facilitate a comprehensive evaluation across varying temporal horizons, we systematically segment and sample the annotated works. Ultimately, our evaluation benchmark comprises a multi-granularity test bed: 200 5-minute, 100 10-minute, 50 15-minute, 40 20-minute, 30 25-minute, and 25 30-minute cinematic clips, specifically designed to stress-test the long-context robustness of multimodal models.

Annotating such dense, multi-modal scripts is notoriously labor-intensive. We adopt an efficient model-in-the-loop paradigm where our automated data engine generates initial dense pseudo-labels, which are subsequently refined by trained annotators. To guarantee utmost reliability, we eschew standard automated consistency checks in favor of a rigorous expert-in-the-loop verification mechanism. Every script undergoes a final comprehensive review by a senior domain expert to cross-reference temporal boundaries and ensure long-term semantic coherence, yielding a benchmark of unprecedented quality.

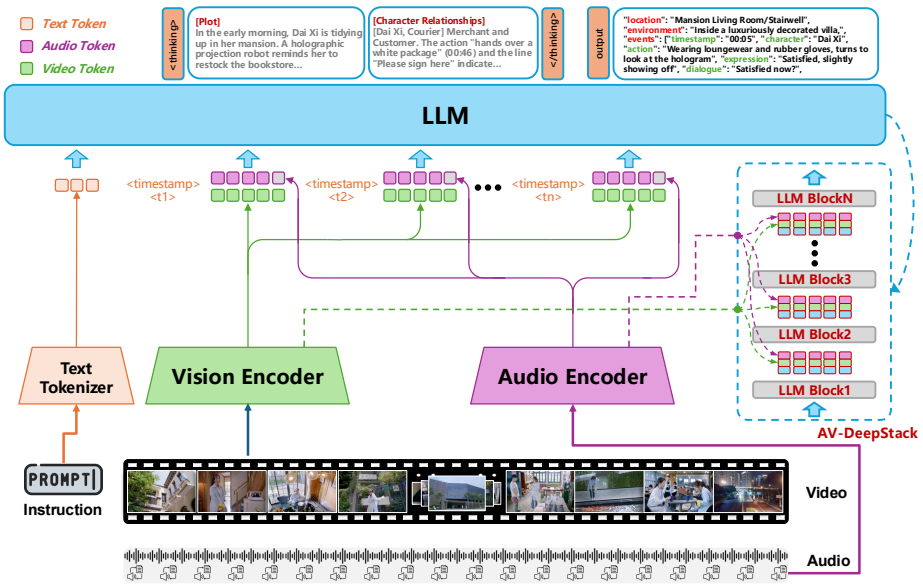

Fig. 3: Overview of the proposed architecture. Instruction, video, and audio are encoded into multimodal tokens and fused in the LLM via AV-DeepStack across multiple layers. The model first performs multimodal plot and character-relationship reasoning and then generates structured script outputs, including location, environment, events, character, action, expression, and dialogue for each event.

5 OmniScript

5.1 Architecture

As shown in Fig. 3, OmniScript follows a unified multimodal reasoning-to-generation pipeline for long-form cinematic video understanding. Given a video clip, the system first encodes visual frames and raw audio into temporally indexed token sequences, then injects these tokens into a large language model (LLM) to perform joint narrative reasoning. The model output is organized into a structured script that explicitly contains scene context (e.g., location and environment) and event-level elements (character, action, expression, dialogue, and audio-aware cues), matching the target schema in Section 3.

Multimodal Temporal Alignment. Beyond conventional vision-only pipelines, we introduce an audio pathway and enforce strict timestamp-level alignment between video and audio features. Specifically, we utilize a pre-trained Whisper [18] encoder to extract audio information. For each temporal unit, the encoder constructs a paired representation (v_t, a_t), where v_t and a_t denote visual and audio embeddings at the same timestamp. This one-to-one alignment preserves cross-modal synchrony for dialogues, off-screen narration, environmental sounds, and background music, which are all critical for script-level narrative grounding.

AV-DeepStack Injection. Our model is built on the Qwen3-VL [3], a vision-language model with native visual-textual reasoning, and adopts a DeepStack-style fusion strategy that injects visual features into multiple LLM layers rather than only at the input stage. Building on this strong VL prior, we extend the architecture to full audio-visual-language modalities. Specifically, after temporal alignment, audio tokens are paired with visual tokens and jointly injected across stacked transformer layers via our AV-DeepStack module. In each layer, the language stream is conditioned by both modalities through residual multimodal adapters, enabling repeated cross-modal interaction during deep semantic inference. This extension preserves the long-context reasoning strengths of the original DeepStack design while adding explicit auditory perception, which is essential for dialogue understanding (including off-screen speech/narration), acoustic event grounding, and BGM-driven emotion interpretation.

Reasoning-Guided Structured Decoding. To maintain global-local consistency, the decoder employs a Chain-of-Thought (CoT) paradigm. Rather than directly predicting event fields, the model first generates an intermediate reasoning trace conditioned on the video segment, comprising (i) a plot progression summary and (ii) an explicit character-relationship state. This trace acts as a structural scaffold for the subsequent coarse-to-fine generation of the temporally grounded script. Specifically, the model establishes the scene context before predicting ordered event records with structured fields (character, action, expression, dialogue, and audio cues). This formulation aligns storyline evolution with event-level details, mitigates long-context ambiguity, and enhances robustness in complex cinematic scenarios involving implicit speaker turns, off-screen audio, and dynamic interpersonal relations.

5.2 Progressive Training and Alignment

OmniScript is optimized via a progressive four-stage pipeline followed by reinforcement learning refinement:

1. Modality Alignment: To integrate the additional audio modality into the vision-language backbone, we first conduct a modality alignment stage. Using approximately 1M bilingual (CN/EN) in-domain cinematic samples with timestamped ASR supervision, we train only the newly introduced audio modules (audio projector) while freezing the original components (Whisper encoder, ViT, and LLM). This parameter-efficient setup preserves pretrained visual-language reasoning while establishing stable cross-modal correspondences. To prevent over-reliance on visual evidence, we randomly mask video frames, encouraging the model to exploit complementary audio cues under partial visual observations.

2. Multimodal Pretraining: Following alignment, we perform large-scale multimodal pretraining to enhance audio-visual-language understanding in long-form cinematic videos. This stage addresses three objectives: (i) unifying cross-modal semantics across diverse narrative styles, (ii) enhancing temporal grounding for event-level localization, and (iii) improving character-centric action and dialogue comprehension. We curate a bilingual corpus of 2.4M in-domain videos and optimize the model using a multi-task objective encompassing ASR (with

and without timestamps), video summarization, dense video captioning, and temporal grounding. Unlike the alignment stage, we fully fine-tune the core components to deeply adapt representations and reasoning layers to the cinematic domain. Random frame masking is retained as a regularization technique.

3. Supervised Fine-Tuning (SFT): We subsequently apply supervised fine-tuning (SFT) to adapt the pretrained model for structured script generation. To improve schema adherence, narrative coherence, and audio-aware descriptions (e.g., BGM and sound effects), we formulate SFT as a Chain-of-Thought (CoT) process. Empirically, plot-conditioned generation yields superior results; thus, we train the fully fine-tuned LLM to first generate an intermediate reasoning trace (capturing plot progression and character relations) before decoding the final structured fields. The SFT dataset comprises 45k in-domain videos (21k horizontal movie/TV clips and 24k vertical short dramas), covering both long cinematic narratives and fast-paced short-video storytelling. The videos are annotated following the pipeline introduced in Section 3.2. Furthermore, we introduce random subtitle masking to reduce reliance on explicit textual cues and improve robustness against missing or noisy subtitles.

4. Reinforcement Learning (RL): To improve the fine-grained descriptive capabilities of the generated scripts, we apply reinforcement learning with verifiable rewards during post-training. Specifically, we optimize the model using GRPO [21] on a small, high-quality dataset of human-annotated scripts. A primary bottleneck in long-sequence, open-ended generation is the design of an effective reward function. Existing methods relying on global semantic similarity often bias toward dominant features, thereby masking subtle errors related to short-duration events. To overcome this, we propose a temporally segmented reward mechanism that evaluates key components across the entire video timeline. The reward is computed using the Multi-dimensional Field Evaluation score (Section 4.1), which performs event-level, one-to-one matching between the generated and ground-truth scripts. This localized alignment allows for the rigorous penalization of fine-grained errors in both recall and precision.

6 Experiments

6.1 Implementation Details

We initialize the model from Qwen3VL-8B and initialize the audio encoder from Whisper large-v3. Our training pipeline follows the three-stage recipe detailed in Sec. 5. In modality alignment, we freeze the LLM backbone and optimize modality projectors to align visual/audio representations with text space, while applying random frame masking to improve robustness under partial observations. In pretraining, we perform full fine-tuning on approximately 2.4M bilingual (Chinese/English) in-domain videos with a unified multi-task objective, including ASR (with/without timestamps), dense captioning, summarization, and temporal grounding. In SFT, we further train on about 45K curated videos (21K movie/TV clips and 24K short-form drama clips) using schema-level supervision

Table 1: Comparison of Event-level metrics for SOTA models on 5-minute videos. Omni indicates whether a model supports audio input (✓: yes, ×: no). In model names, -T denotes Thinking mode. For MoE models, parameters are reported as total parameters/activated parameters.

Model	Param./B	Omni	Char.	Dia.	Act.	Exp.	Aud.	Overall	tIoU@0.1
Proprietary Models									
Gemini-3-flash		✓	28.8	50.3	28.2	25.5	11.2	28.8	44.3
Gemini-3-pro		✓	39.8	68.8	37.4	35.4	13.3	38.9	64.4
Gemini-2.5-flash		✓	40.1	75.5	42.8	36.5	22.8	43.6	74.3
Gemini-2.5-pro		✓	41.7	75.0	41.9	39.0	17.0	42.9	73.4
Seed-1.8		✓	40.9	54.4	35.1	29.6	12.4	34.5	50.7
Seed-2.0-pro		✓	47.4	68.1	42.9	35.7	10.3	40.9	67.1
Open-source Models									
Qwen3VL	8	×	30.4	49.6	26.9	25.3	6.6	27.7	47.6
Qwen3-Omni-T	30/3	✓	3.2	3.5	3.8	6.8	2.3	3.9	3.0
Qwen3-Omni	30/3	✓	4.9	3.4	5.5	7.3	4.4	5.1	12.8
Qwen3VL-T	32	×	24.4	37.5	22.1	18.9	7.0	22.0	34.6
Qwen3VL	32	×	37.1	57.1	31.3	28.7	7.2	32.3	52.5
Qwen3VL-T	235/22	×	35.7	57.6	27.4	23.9	6.5	30.2	54.8
Qwen3VL	235/22	×	38.1	58.6	33.0	29.1	6.0	33.0	62.0
Ours	8	✓	39.2	72.2	33.7	31.9	11.6	37.7	69.3

with CoT-style intermediate traces (plot evolution and character relation reasoning) before final structured decoding. All performance results and comparisons in this section are conducted on our proposed benchmark.

6.2 Performance comparison on 5-Minute Videos

Tables 1 and 2 compare our model with representative proprietary and open-source models under a unified protocol. We report F1 scores for all event/scene fields, and *Overall* is their mean, reflecting the accuracy of event content understanding. For temporal localization, we use tIoU@0.1. A key takeaway is parameter efficiency: our model uses only 8B parameters, yet delivers strong performance on both event content quality and temporal localization.

Event-level comparison. In Table 1, our 8B model achieves 37.7 Overall score and 69.3 tIoU@0.1. It substantially outperforms open-source models of much larger scales, gaining +4.7 in Overall score and +7.3 in tIoU@0.1 compared to Qwen3VL-235B-A22B. Against proprietary models, it achieves stronger dialogue understanding than Gemini-3-pro [1] and Seed-2.0-pro [2], and also surpasses them in temporal localization. We also observe that thinking variants are often worse than their non-thinking counterparts, and existing omni-input variants (e.g., Qwen3-Omni) are notably weaker than similarly sized non-omni models.

Scene-level comparison. In Table 2, our model reaches a 52.4 Overall score and 74.6 tIoU@0.1. Relative to Qwen3VL-235/22B, it shows better temporal boundary quality (+1.8 on tIoU@0.1) at a much smaller scale. Scene attributes

Table 2: Comparison of Scene-level metrics for SOTA models on 5-minute videos. Omni indicates whether a model supports audio input (✓: yes, ×: no).

Model	Param./B	Omni	Loc.	Type	Env.	Time	Mood	Overall	tIoU@0.1
Proprietary Models									
Gemini-3-flash		✓	54.6	59.8	42.7	54.9	50.4	52.5	70.3
Gemini-3-pro		✓	58.8	63.1	46.9	61.6	54.8	57.0	75.3
Gemini-2.5-flash		✓	52.8	57.1	45.7	56.1	50.3	52.3	69.6
Gemini-2.5-pro		✓	56.6	62.4	50.8	60.1	54.6	56.9	74.1
Seed-1.8		✓	57.9	58.6	47.7	58.7	52.8	55.1	74.0
Seed-2.0-pro		✓	57.7	62.2	49.2	62.7	54.3	57.2	75.5
Open-source Models									
Qwen3VL	8	×	41.3	49.7	31.8	39.8	41.7	40.9	60.6
Qwen3-Omni	30/3	✓	18.4	26.0	14.6	23.6	22.4	21.0	29.6
Qwen3VL	32	×	50.4	58.7	42.7	55.4	47.9	51.0	71.1
Qwen3VL	235/22	×	52.6	60.2	45.4	57.9	50.9	53.4	72.8
Ours	8	✓	54.0	58.4	41.9	58.1	49.5	52.4	74.6

requiring subtle visual-audio context integration, such as environment and mood, remain challenging across all evaluated models.

6.3 Performance Comparison on Longer Videos

To evaluate the robustness of multimodal models in processing extended temporal contexts, we further benchmark performance across varying video durations, extending up to 30-minute cinematic segments. As illustrated in Fig. 4, several critical insights emerge from this challenging setting.

The Challenge of Long-Form Context. There is a clear negative correlation between video duration and the Overall F1 score across all evaluated models. The degrading trend highlights the inherent difficulty of long-form cinematic understanding, where models suffer from error accumulation, context dilution, and the vanishing of fine-grained details when overwhelmed by massive streams of audio-visual tokens.

Superior Resilience of Our Framework. Despite the general downward trend, the performance decay of our model is comparatively more gradual. Notably, on the exceptionally challenging 30-minute video subset, our model surpasses the highly competitive `seed-2.0-pro`. The sustained superiority demonstrates that our architectural design mitigate catastrophic forgetting and maintains narrative consistency over long horizons.

The Context-Capacity Trade-off in Gemini-2.5-Flash. An intriguing phenomenon is the dichotomous behavior of gemini-2.5-flash. For videos up to 20 minutes, it comprehensively scans and reports dense events without severe omission, resulting in exceptionally high recall. However, as duration extends beyond the 20-minute threshold, its accuracy experiences a precipitous drop. Constrained by its model capacity, gemini-2.5-flash struggles to adhere to the rigid hierarchical script formatting over extreme lengths, frequently degenerating into

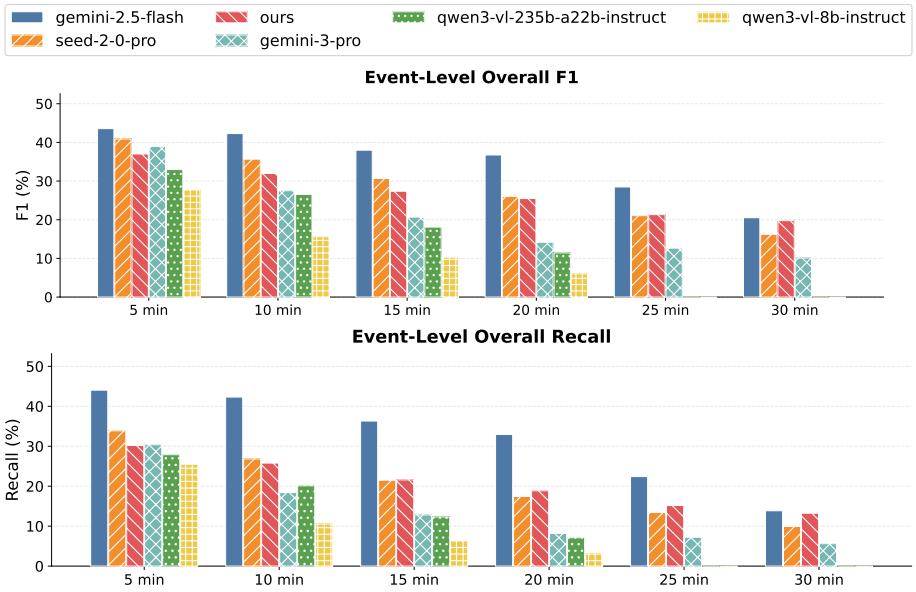

Fig. 4: Comparison of Overall F1 and recall score on different duration videos.

Table 3: Ablation study for training strategy.

Model	CoT	Reward	Char.	Dia.	Act.	Exp.	Aud.	Overall	tIoU@0.1
SFT	×	–	35.6	68.2	30.5	31.2	11.12	35.3	66.6
SFT	✓	–	37.8	71.0	33.5	31.2	11.5	37.0	68.9
SFT+RL	×	Segmented	37.1	70.9	32.8	32.5	11.7	37.0	69.0
SFT+RL	✓	Global	39.2	69.0	32.4	31.8	12.3	37.0	68.7
SFT+RL	✓	Segmented	39.2	72.2	33.7	31.9	11.6	37.7	69.3

repetitive generation loops and unstructured outputs. This underscores that a massive theoretical context window does not necessarily equate to sustained structural instruction-following capability

6.4 Ablation Studies

Effectiveness of training strategy. Table 3 ablates our Chain-of-Thought decoding and Reinforcement Learning (RL) stages. **Reasoning Trace (CoT):** Introducing CoT to the SFT baseline boosts the Overall score (35.3% → 37.0%) and Dialogue F1 score (68.2% → 71.0%). By constructing a latent cognitive scaffold prior to event decoding, CoT successfully enforces global-local consistency and significantly reduces long-context ambiguity. **RL Alignment:** Transitioning to the RL stage further increases the Overall score to 37.7%. By optimizing sequence-level metrics, RL effectively mitigates the exposure bias inherent in

Table 4: Comparison of event-level performance with and without masking video subtitles. SV denotes Subtitle Visible (✓: visible, ×: masked).

Model	SV	Char.	Dialog	Action	Exp.	Audio	Overall	tIoU@0.1
Qwen3VL-235B-A22B	✓	38.1	58.6	33.0	29.1	6.0	33.0	62.0
Qwen3VL-235B-A22B	×	26.2	7.7	29.8	23.5	6.0	18.6	45.1
Gemini-3-pro	✓	39.8	68.8	37.4	35.4	13.3	38.9	64.4
Gemini-3-pro	×	40.4	60.9	34.7	33.6	13.2	36.6	60.3
Ours-8B	✓	39.2	72.2	33.7	31.9	11.6	37.7	69.3
Ours-8B	×	34.1	63.8	31.8	30.6	11.7	34.4	67.0

Table 5: Comparison of event-level performance with and without audio as input.

Model	Omni	Char.	Dia.	Act.	Exp.	Aud.	Overall	tIoU@0.1
Qwen3-VL-8B-SFT	×	34.0	52.0	30.3	28.5	10.5	31.1	68.2
Ours-8B-SFT	✓	35.6	68.2	30.5	31.2	11.1	35.3	66.6

SFT auto-regressive decoding. **Segmented Reward:** Crucially, our proposed Segmented Reward outperforms the standard Global reward by rigorously penalizing fine-grained precision and recall errors at the event level.

Effectiveness of video subtitle. In Tables 1 and 2, several non-omni models (without audio input) still obtain relatively strong performance on dialogue F1 score, suggesting a potential shortcut from visible subtitles. To diagnose this effect, we mask subtitles at inference time and report the comparison in Table 4. After subtitle masking, omni models such as Gemini-3-pro show a moderate degradation on dialogue F1 score (68.8 → 60.9), indicating that dialogue recognition is not purely text-copying and still relies on multimodal cues. This also suggests that omni models are more robust in subtitle-free settings. In contrast, Qwen3VL-235B-A22B exhibits a severe collapse (58.6 → 7.7), which suggests substantial dependence on visual information rather than robust audio-visual dialogue understanding.

Effectiveness of Audio-Injection Pretraining. To rigorously validate the necessity of the acoustic modality and our pretraining stage, we conduct an ablation study comparing our full model against a vision-only baseline. Specifically, we train a `Qwen3-VL-8B-SFT` model directly on our constructed dataset using only visual frames, completely depriving it of audio inputs. As illustrated in Table 5, while the vision-only baseline manages to capture basic visual actions (scoring 30.3% in Act.), it suffers a significant performance bottleneck in audio-dependent and implicit semantic fields. Most notably, the Dialogue recognition accuracy of the vision-only model is constrained to 52.0%. In stark contrast, our full model, which leverages the audio-injection pretraining phase to align acoustic features with the LLM backbone, achieves a remarkable 68.2% in Dialogue accuracy, yielding a massive absolute improvement of +16.2%. This substantial

performance gap underscores the fundamental limitation of vision-centric models in cinematic scripting and validates the effectiveness of our pretraining stage.

7 Conclusion

In this paper, we present *OmniScript*, an audio-visual language model for script-oriented understanding of long-form narrative videos. We target two complementary goals: accurate multi-field semantic parsing (event and scene attributes) and reliable temporal localization. We build the first comprehensive video script benchmarks with high-quality manual annotations, featuring long, complex cinematic videos. Extensive experiments on our benchmark show that OmniScript achieves a strong balance between quality and efficiency. With only 8B parameters, it consistently outperforms much larger open-source models on event-level understanding and temporal grounding, while remaining competitive on scene-level metrics. The subtitle-masking analysis further highlights robustness differences between model families and confirms the value of genuine audio-visual reasoning beyond text shortcuts. Our findings suggest that robust script understanding still depends on better fine-grained multimodal perception, especially for subtle attributes and boundary-sensitive localization. We hope this work provides a useful baseline, benchmark, and set of training insights for future research on omni-modal narrative video understanding.

References

1. Gemini 3 pro model card (2025), `https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf`, last accessed 2026/06/30
2. Seed 2.0 model card (2026), `https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf`, last accessed 2026/06/30
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-VL technical report. arXiv preprint arXiv:2511.21631 (2025)
4. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: Proceedings of the Asian Conference on Computer Vision (2020)
5. Chen, X., Ding, Y., Lin, W., Hua, J., Yao, L., Shi, Y., Li, B., Zhang, Y., Liu, Q., Wan, P., et al.: AVoCaDO: An audiovisual video captioner driven by temporal orchestration. arXiv preprint arXiv:2510.10395 (2025)
6. Chen, X., Lin, W., Hua, J., Yao, L., Ding, Y., Li, B., Zeng, B., Shi, Y., Liu, Q., Zhang, Y., et al.: DiaDem: Advancing dialogue descriptions in audiovisual video captioning for multimodal large language models. arXiv preprint arXiv:2601.19267 (2026)
7. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)

8. Ge, Y., Ge, Y., Li, C., Wang, T., Pu, J., Li, Y., Qiu, L., Ma, J., Duan, L., Zuo, X., et al.: ARC-Hunyuan-Video-7B: Structured video comprehension of real-world shorts. arXiv preprint arXiv:2507.20939 (2025)
9. Geng, T., Zhang, J., Wang, Q., Wang, T., Duan, J., Zheng, F.: LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18959–18969 (2025)
10. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6047–6056 (2018)
11. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: MovieNet: A holistic dataset for movie understanding. In: European conference on computer vision. pp. 709–727. Springer (2020)
12. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: Proceedings of the IEEE international conference on computer vision. pp. 706–715 (2017)
13. Li, C., Peng, X., Wang, T., Ge, Y., Liu, M., Xu, X., Wang, Y., Shan, Y.: PTVD: A large-scale plot-oriented multimodal dataset based on television dramas. arXiv preprint arXiv:2306.14644 (2023)
14. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
17. Pu, J., Wang, T., Ge, Y., Ge, Y., Li, C., Shan, Y.: ARC-Chapter: Structuring hour-long videos into navigable chapters and hierarchical summaries. arXiv preprint arXiv:2511.14349 (2025)
18. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: ICML. pp. 28492–28518 (2023)
19. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015)
20. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123**(1), 94–120 (2017)
21. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
22. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., et al.: LongVU: Spatiotemporal adaptive compression for long video-language understanding. arXiv preprint arXiv:2410.17434 (2024)
23. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5026–5035 (2022)

24. Tang, C., Li, Y., Yang, Y., Zhuang, J., Sun, G., Li, W., Ma, Z., Zhang, C.: video-SALMONN 2: Caption-enhanced audio-visual large language models. arXiv preprint arXiv:2506.15220 (2025)
25. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
26. Torabi, A., Pal, C., Larochelle, H., Courville, A.: Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070 (2015)
27. Vicol, P., Tapaswi, M., Castrejon, L., Fidler, S.: MovieGraphs: Towards understanding human-centric situations from videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8581–8590 (2018)
28. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
29. Yao, L., Wei, Y., Zhang, Y., Li, L., Chen, X., Song, F., Wang, Z., Ouyang, K., Liu, Y., Kong, L., et al.: TimeChat-Captioner: Scripting multi-scene videos with time-aware and structural audio-visual captions. arXiv preprint arXiv:2602.08711 (2026)
30. Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., Cai, T., Chen, W., Huang, Y., Zhao, Y., et al.: MiniCPM-V 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025)
31. Yue, Z., Zhang, Q., Hu, A., Zhang, L., Wang, Z., Jin, Q.: Movie101: A new movie understanding benchmark. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4669–4684 (2023)
32. Yue, Z., Zhang, Y., Wang, Z., Jin, Q.: Movie101v2: Improved movie narration benchmark. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 17081–17095 (2025)
33. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: LLaVA-Video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
34. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)