Movie101v2: Improved Movie Narration Benchmark

Zihao Yue[*], **Yepeng Zhang**[*], **Ziheng Wang**, **Qin Jin**
Renmin University of China
{yzihao, yepzhang, zihengwang, qjin}@ruc.edu.cn
https://movie101-dataset.github.io

Abstract

Automatic movie narration aims to generate video-aligned plot descriptions to assist visually impaired audiences. Unlike standard video captioning, it involves not only describing key visual details but also inferring plots that unfold across multiple movie shots, presenting distinct and complex challenges. To advance this field, we introduce **Movie101v2**, a large-scale, bilingual dataset with enhanced data quality specifically designed for movie narration. Revisiting the task, we propose breaking down the ultimate goal of automatic movie narration into three progressive stages, offering a clear roadmap with corresponding evaluation metrics. Based on our new benchmark, we baseline a range of large vision-language models, including GPT-4V, and conduct an in-depth analysis of the challenges in narration generation. Our findings highlight that achieving applicable movie narration generation is a fascinating goal that requires significant research.

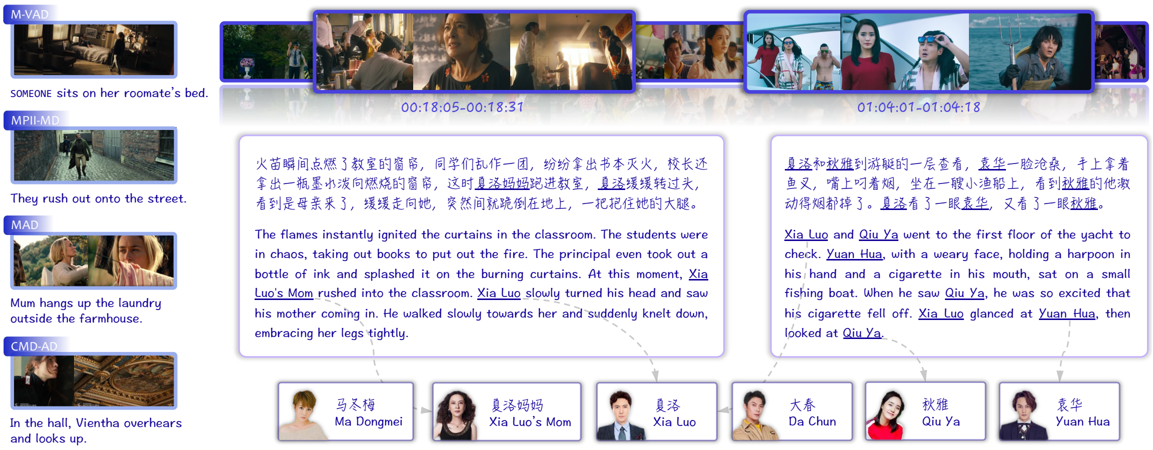
1 Introduction

Audio Description (AD) is a crucial technology that enables visually impaired individuals to enjoy movies by integrating real-time voice narrations into the movie soundtrack, describing the movie's visual content. Unlike video captions, movie narrations must briefly summarize the ongoing content during pauses in character dialogue, providing accurate descriptions of visual facts (e.g., scenes, characters, and events) and key plot points to help the audience stay engaged. However, creating movie narrations usually requires extensive manual effort from trained professionals, making it impractical to cover the vast number of movies and TV shows online. Therefore, researchers have begun exploring automatic movie narration generation, advancing research from **data** (Torabi et al., 2015; Soldan et al., 2022), **task** (Rohrbach et al., 2017; Han et al., 2023b), and **method** (Han et al., 2023a; Yue et al., 2023) perspectives, but there is still a long way to go before realizing the ultimate goal of automatically generating high-quality and applicable movie narrations.

From the data perspective, proper benchmark datasets play a fundamental role in the development of movie understanding and narrating. Early efforts primarily focus on basic video-to-text descriptions with simplified narration data (Rohrbach et al., 2017; Soldan et al., 2022). For example, LSMDC (Rohrbach et al., 2017) replaces character names in narrations with "someone", reducing the movie narration task to a general video captioning task. AutoAD (Han et al., 2023b) improves on this by reinstating character names from the raw data of MAD (Soldan et al., 2022), while AutoAD II (Han et al., 2023a) goes further by introducing a character bank to support generating narrations with character names. However, the movie clips designated for narration in these datasets are typically short (e.g., on average 6.2 seconds in M-VAD (Torabi et al., 2015) and 4.1 seconds in MAD), as shown in Fig. 1, which limits the ability of models to generate coherent narrations for longer clips based on complex visual changes and plot sequences.

The recently proposed dataset, Movie101 (Yue et al., 2023), addresses some of these limitations by providing longer video narration paragraphs and rich movie metadata, such as character information. However, our analysis identifies several drawbacks. First, Movie101 consists of only 101 movies with 14,000 video-narration pairs, which is smaller in scale compared to other available datasets. Second, it includes only Chinese narrations, limiting its usefulness for non-Chinese speakers and hindering the application of many advanced English-based models for movie understanding. Third, the automatically crawled metadata contains errors, such as missing characters in cast lists and inconsistent character names in narrations. To address these

[*]Equal contribution.



SOMEONE sits on her roomate's bed.

They rush out onto the street.

Mum hangs up the laundry outside the farmhouse.

In the hall, Vientha overhears and looks up.

火苗瞬间点燃了教室的窗帘，同学们乱作一团，纷纷拿出书本灭火，校长还拿出一瓶墨水泼向燃烧的窗帘，这时夏洛妈妈跑进教室，夏洛缓缓转过头，看到是母亲来了，缓缓走向她，突然间就跪倒在地上，一把抱住她的大腿。

The flames instantly ignited the curtains in the classroom. The students were in chaos, taking out books to put out the fire. The principal even took out a bottle of ink and splashed it on the burning curtains. At this moment, Xia Luo's Mom rushed into the classroom. Xia Luo slowly turned his head and saw his mother coming in. He walked slowly towards her and suddenly knelt down, embracing her legs tightly.

夏洛和秋雅到游艇的一层查看，袁华一脸沧桑，手上拿着鱼叉，嘴上叼着烟，坐在一艘小渔船上，看到秋雅的他激动得烟都掉了。夏洛看了一眼袁华，又看了一眼秋雅。

Xia Luo and Qiu Ya went to the first floor of the yacht to check. Yuan Hua, with a weary face, holding a harpoon in his hand and a cigarette in his mouth, sat on a small fishing boat. When he saw Qiu Ya, he was so excited that his cigarette fell off. Xia Luo glanced at Yuan Hua, then looked at Qiu Ya.

Figure 1: Examples from other datasets (left) and Movie101v2 (right) where cases are from *Goodbye Mr. Loser*.

issues and to facilitate automatic movie narration from a data perspective, we improve Movie101 in terms of both *scale* and *quality*. By collaborating with various expert models, we manage to avoid the previously used heavy crowd-sourcing and increase the dataset cost-effectively to a scale of 203 movies and 46,000 bilingual (Chinese and English) video-narration pairs, which we name **Movie101v2**.

From the task perspective, the development of automatic movie narration has evolved over time. AutoAD highlights the dependence of movie narration on context, incorporating a preceding narration and movie subtitles as additional task inputs. Movie101 and AutoAD II address practical deployment issues of *where* to insert narrations, by narrating between character dialogues or incorporating a timestamp prediction task. While these task definitions strive for more applicable narrations in real-world settings, our closer inspection of the Movie101v2 data reveals that achieving fully deployable movie narrations remains an ambitious goal.

On the one hand, generating applicable narrations requires handling complex inputs like extensive plot histories and character dialogues, which are beyond the current capabilities of most models. On the other hand, our initial experimental findings suggest that even understanding the basic visual facts and movie plots within *individual* movie clips is a fundamental but unsolved challenge. Therefore, the ambitious goal of achieving applicable movie narrations will require incremental progress to achieve step by step. To address this, we suggest breaking down the long-term goal into three progressively challenging stages: basic visual facts description (L1), plot narration (L2), and applicable AD text generation (L3). Given the current technological limitations, we prioritize achieving the L1 and L2 goals, focusing on movie understanding within individual clips.

To align with these staged goals, the evaluation framework should also evolve to provide more targeted feedback for model improvement. Existing works typically assess the linguistic or semantic matching between generated narrations and reference ones (ground truth) (Rohrbach et al., 2017; Han et al., 2023b,a; Yue et al., 2023). However, since reference narrations are produced by human experts with access to extensive contextual information (e.g., prior plot developments and character histories), directly comparing them to model outputs, which are generated based solely on the given video clip, may not be a fair evaluation. To address this, we propose a new evaluation framework that leverages Large Language Models (LLMs) to separately assess narrative quality from the L1 and L2 perspectives. This helps avoid comparing against out-of-context information, and offers a clearer guidance for model development and improvement.

From the method perspective, we explore practical solutions for movie narration generation. We build baselines based on several state-of-the-art Large Vision-Language Models (LVLMs), including both open-source models and GPT-4V (OpenAI, 2023), on Movie101v2 in both Chinese and English. This provides a comprehensive evaluation of their performance in movie narration generation. We also carry out detailed analytical experiments to identify the challenges and difficulties that current models face, both in terms of visual perception and text generation. These findings aim to provide practical insights and inspire future research into improving automatic movie narration generation.

In summary, in our pursuit of advancing movie narration generation, large-scale and high-quality

Table 1: Dataset statistics. $L_{\mathrm{v/t}}$: average video duration (sec.) or text length (en. words or zh. characters); N_{char}: character count; Gray: datasets with short movie clips.

Dataset	# movie	# clip	L_{v}	L_{t}	N_{char}
M-VAD	92	49K	6.2	9.1	-
MPII-MD	94	68K	3.9	9.6	-
MAD	650	385K	4.1	12.7	-
CMD-AD	1,432	101K	-	-	-
Movie101 (zh)	101	14K	20.4	80.7	2.0
Movie101v2-zh	203	46K	12.8	60.0	1.9
Movie101v2-en	203	46K	12.8	39.1	1.9

data are the cornerstone, clearly defined tasks and evaluations propel progress, and benchmarking and analysis of advanced models share important findings. Through the development of Movie101v2 in this work, our contributions in data, task design, evaluation framework, and experimental insights aim to shed some light on further research and progress in movie narration generation.

2 Data: Movie101v2

The original Movie101 dataset contains 101 movies from the barrier-free channel of Xigua Video[1], where movies are reproduced with video-aligned narration speeches. The narration texts are obtained through ASR transcription and refined with crowd-sourcing. In this section, we improve upon Movie101 in both data scale and quality.

2.1 Scaling Up

Following Movie101, we collect all the newly available movies with narrations from Xigua Video (102 new movies in total), obtaining narration data through the following automatic process.

Speech Transcription. We transcribe movie audio into text by Whisper (OpenAI, 2022). The raw ASR output includes both narrations and character dialogues, and often contains transcription errors.

Text Refinement. To remove character dialogues from the ASR outputs, we first use OCR (PaddleOCR, 2022) to detect movie subtitles, thereby identifying the instances where dialogues occur and subsequently removing them. This process concurrently produces subtitle data equipped with precise timestamps. Next, we use GPT-4 to identify and remove any remaining dialogues. Upon filtering out dialogues, we leverage GPT-3.5-turbo to correct textual errors within the narrations, including typos, wrong punctuation, and nonsensical phrases.

Clip Merging. To support comprehension of coherent plots conveyed in consecutive movie clips, we merge adjacent narration segments into paragraphs. Employing a heuristic approach, we utilize a dynamically adjusted threshold to determine if two adjacent segments should be merged, thereby ensuring effective merging near clips while preventing the creation of excessively long paragraphs.

Translation. Following data refinement, we translate Chinese narrations into English using GPT-3.5-turbo. To ensure accurate translation of character names within narrations, we manually construct an English version of the movie casts, serving as references for the LLM.

2.2 Quality Enhancement

Character Name Refinement. We collect movie metadata, including basic movie information and cast details, from Xigua Video. However, our manual review reveals two significant issues not identified in Movie101: (1) The cast list for some movies is incomplete, with some or all character names missing. (2) The character names in narrations often do not match their official cast names. Additionally, the same character may be narrated by different names throughout a movie, further limiting the connection between narrated characters and external cast knowledge. To address these issues, we refine character names in both the movie casts and narrations for Movie101 and our newly collected data. First, we complete and correct names in the cast lists through human annotation. Then, we use GPT-3.5-turbo to automatically update the narrations, ensuring the character names in the narrations precisely match those in the movie casts.

Quality Control. While our data construction leverages LLMs to minimize labor costs, it is crucial to ensure and monitor the quality of the automatically refined narrations. To maintain data quality, we have LLMs focus on one refining step at a time, even though they are capable of handling multiple steps (e.g., dialogue removal and textual correction) in a single response. Additionally, recognizing the importance of contextual clues, such as the co-occurrence of characters and objects in adjacent clips, we implement a batching strategy that allows LLMs to reference surrounding contexts when refining narrations. To further enhance performance, we provide input-output demonstrations for each task, taking advantage of the in-context learning capabilities of LLMs. A manual review of 300 narration samples (see Appendix A) shows

[1] https://www.ixigua.com/channel/barrier_free

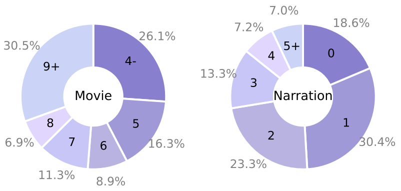


Figure 2: Distribution of character counts in movie casts (left) and narration paragraphs (right).

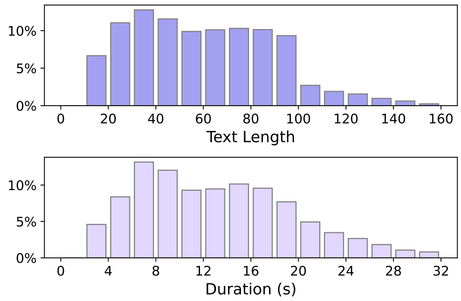


Figure 3: Distribution of narration text length and movie clip duration in Movie101v2-zh.

that Movie101v2 exhibits competitive data quality compared to human-refined Movie101.

2.3 Data Statistics

Movies. Movie101v2 comprises 203 movies totaling 353 hours. On average, each movie features 7.3 characters, with the distribution of character counts detailed in Fig. 2. The most popular genres include comedy, romance, and action. The dataset follows the same split as Movie101, with 10 movies each allocated for validation and testing.

Narrations. We collect 71K video-aligned narration segments and merge them into 46K narration paragraphs, with the length distribution shown in Fig. 3. Detailed statistics comparing Movie101v2 to other datasets are available in Table 1. Unlike Movie101, which defines the entire interval between two character dialogues as a single clip for narrating, often resulting in sparse narration annotations over a long clip, Movie101v2 provides denser and more closely video-aligned narrations within corresponding clips.

Discussion. By leveraging expert models and LLMs, we establish a streamlined, repeatable, and cost-friendly process for acquiring narration data, laying the groundwork for future dataset expansion. Beyond increasing the data scale, we also enhance the character information to better link narrations with external character knowledge. We hope this enriched data will support the research community in future studies on movie narrating and related tasks, such as temporal grounding in movies (Gao et al., 2017; Yue et al., 2023), visual question answering (Song et al., 2023), and more.

3 Task: Movie Narration Generation

This section revisits the task of automatic movie narration generation. We break down the long-term goal of movie narration into three progressive goals (Section 3.1), and present a new evaluation framework for narration quality assessment (Section 3.2).

3.1 Task Roadmap

Movie narrations differ from standard video descriptions in that they must not only provide accurate and focused details of visual facts but also infer the plot of the movie by combining information from multiple shots. Effectively narrating a given clip may also rely on the plot's history, as well as cues from sound, character dialogues, and other elements. To generate high-quality, deployable movie narrations automatically, current studies have explored a range of task inputs and outputs. For example, AutoAD highlights the need for various contextual inputs in movie narration generation, using a preceding narration and movie subtitles as additional inputs. However, relying on neighbor contexts alone is often insufficient to capture long-term plot history, and depending on previous narrations can lead to cumulative errors. Producing context-aware narrations may demand complex and extensive inputs, potentially exceeding the capacity of current models. Both Movie101 and AutoAD II consider the timing of narration generation for practical deployment, with the latter requiring explicit prediction of narration timestamps. However, as our preliminary experiments show, current models struggle with even the most basic narration generation task, suggesting that additional requirements may be unnecessary distractions at this stage.

Given the current state of model development, achieving perfect movie narration remains a challenging, long-term goal. To create a clear roadmap to facilitate developing effective movie narration systems, we propose breaking down this ultimate goal into three progressive stages:

L1. Visual Facts Description. This stage focuses on providing accurate and comprehensive descriptions of key visual facts in a given movie clip.

L2. Plot Narration. This stage involves reasoning across multiple shots to describe the plot of the current clip. We emphasize that this stage goes beyond L1, as movies convey plots through the sequence

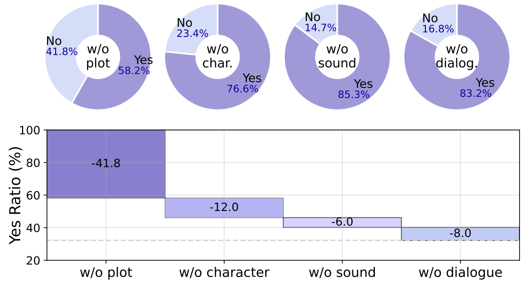


Figure 4: Human evaluation results on the necessity of various multi-modal contextual information for L3 narrations. The top section displays the percentage of "Yes" responses for whether a clip can be accurately narrated without corresponding context. The bottom section shows the "Yes" percentage with all context information being removed progressively.

of shots, relying on human cognitive abilities to piece together information fragments into a coherent story. Such a sense of story cannot be achieved by simply listing visual facts from each shot. We provide examples in Appendix B to illustrate the distinction between L1 and L2.

L3. Applicable Movie AD. This final stage aims to produce high-quality narration scripts that are not only accurate and coherent but also appropriately timed and paced, making them directly applicable for practical deployment.

The goals of L1 and L2 focus on understanding what occurs within a specific movie clip, i.e., the visual facts and local plots. In contrast, L3 necessitates a more comprehensive understanding, integrating not just individual clips but also the multi-modal context of the entire movie. To achieve the ultimate goal of L3 step by step, we expect first to achieve the foundational goals of L1 and L2, concentrating on understanding and narrating the content of individual clips.

3.2 Evaluation

Existing works primarily evaluate the quality of the generated narrations by comparing them to the reference ones (ground truth), using matching-based metrics such as CIDEr (Vedantam et al., 2015), BLEU (Papineni et al., 2002), and ROUGE (Lin, 2004), feature-based metrics such as BERTScore (Zhang et al., 2019), and LLM-based metrics (Han et al., 2024). However, the reference narrations are derived from deployable ADs created by human experts based on a wealth of contextual information, which is not available to models in the L1 and L2 task settings. This creates a clear mismatch between the reference narrations and the task goals, raising concerns that direct similarity measurements may not provide an appropriate assessment. To investigate how the "missing contexts" affect narration generation, including plot histories, sound, dialogues, and character context (characters are easier to identify given prior appearances), we conduct a human evaluation. We randomly sample 1,000 movie clips from the 10 test set movies, and ask annotators whether specific context information is necessary to generate the reference-quality narration.

As shown in Fig. 4 (top), generating reference-quality (or L3) narrations often requires multi-modal context information beyond visual inputs, particularly the plot history and character context. By incrementally removing these contexts (Fig. 4 (bottom)), the ability to generate accurate narrations significantly declines. This highlights the challenge of achieving perfect movie narrations, and supports our argument that reference narrations are often unattainable due to limited model inputs. Consequently, evaluating narrations solely based on their similarity to reference narrations can be misleading and may fail to provide useful feedback for optimizing current narration models.

To improve the evaluation process, we introduce two new metrics aimed at separately assessing narration quality at the L1 and L2 levels:

L1-Score evaluates how well the generated narration describes visual facts present in the reference, focusing on (a) *Environment*, including scenes, objects, and events; (b) *Character*, including character names, actions, and emotions.

L2-Score measures how consistently the generated narration conveys the plot compared to the reference, regardless of the linguistic similarity and specific visual details.

For each metric, LLM rates on a scale from 0 to 5. These metrics provide a more comprehensive evaluation of model performance, better aligning with our L1 and L2 task goals.

4 Method: Benchmarking and Analysis

4.1 Benchmarking

Large Vision-Language Models (LVLMs) have recently become the leading approach for video-to-text tasks. We benchmark state-of-the-art LVLMs, including VideoGPT+ (Maaz et al., 2024), VideoChat-2 (Li et al., 2023b), VideoLLaMA 2 (Cheng et al., 2024), InternVL2 (Chen

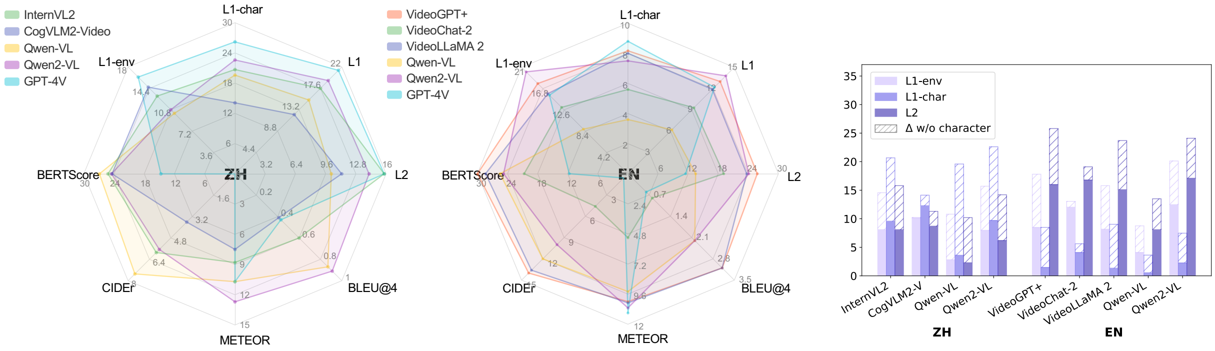


Figure 5: Model performance on Movie101v2. All L1/L2 scores are rescaled to a range of 0-100.

et al., 2023b), CogVLM2-Video (Hong et al., 2024), Qwen-VL (Bai et al., 2023), Qwen2-VL (Wang et al., 2024), and GPT-4V, on our movie narration task, which can be broadly categorized into two groups: models that process videos directly and models that process multiple images. The task input consists of a video V comprising L frames $\{f_1, \ldots, f_L\}$, character portraits $C_{\text{portrait}} = \{p_1, \ldots, p_m\}$ and corresponding character names $C_{\text{name}} = \{n_1, \ldots, n_m\}$. The goal is to generate a narration $\hat{y}$ for V. Below, we outline how we adapt the two types of models for this task:

Video Models typically include a video encoder $\mathcal{F}(\cdot)$ to extract video features and a projector $\mathcal{P}(\cdot)$ to convert these features into visual tokens. The LLM then generates narrations based on the visual tokens and task instructions I. To incorporate character information, we treat C_{portrait} as additional video frames, and provide C_{name} as textual inputs of the LLM. The data flow is defined as:

$$\hat{y} = \text{LLM}\Big(\mathcal{P}\big(\mathcal{F}(V; C_{\text{portrait}})\big); I; C_{\text{name}}\Big),$$

where ";" denotes sequence concatenation. We avoid encoding videos and portraits independently with $\mathcal{P}(\cdot)$, as this can cause the resulting visual tokens to lose critical facial information. Instead, we fuse their respective features early, before projection, which has been shown to improve performance in our preliminary experiments.

Multi-image Models do not explicitly support video input, but can perform inference across multiple video frames. Ideally, we could provide each frame and portrait to the model. However, due to the limited context length (typically supporting up to K images), we divide the video into K segments and concatenate adjacent frames into a single image, resulting in $V' = \{v'_i\}_{i=1}^{K}$. Similarly, we concatenate character portraits into p'. To better associate portraits with character names, we also add visual text to the portrait images. The overall process is defined as:

$$\hat{y} = \text{LLM}\Big(\big\{\mathcal{P}\big(\mathcal{F}(v'_i)\big) | v'_i \in V'\big\}; \mathcal{P}\big(\mathcal{F}(p')\big); I; C_{\text{name}}\Big).$$

We finetune open-sourced models on the Movie101v2 training set for 3 epochs, freezing the vision encoder and training only the visual projector and LoRA (Hu et al., 2022) adapters of the LLM. For GPT-4V, which cannot be finetuned with our data, we provide a carefully designed task input to encourage in-context learning. This input includes a detailed task description and several randomly retrieved narration examples from the training set to guide the model toward generating narrations in the appropriate style. Implementation details are provided in Appendix C.1.

Evaluation. The L1/L2 scores of Chinese and English narrations are obtained using DeepSeek-V2.5 (DeepSeek-AI, 2024) and Llama-3.1-70B-Instruct (Dubey et al., 2024), respectively. We use these open-sourced models to ensure the best reproducibility, as commercial LLM APIs can be discontinued or updated. Additionally, evaluation results obtained using GPT-3.5-turbo are provided in Appendix D.1. Along with the L1/L2 scores, we also report performance on BLEU, METEOR, CIDEr, and BERTScore for further reference.

Results. Fig. 5 shows the performance of various models on Movie101v2. The multi-image model GPT-4V serves as a strong baseline, particularly in the Chinese setting, despite not being finetuned specifically for this task. Among open-sourced models, VideoGPT+, VideoLLaMA 2, InternVL2, and Qwen2-VL show comparable performance, with some excelling in L1 (visual facts) and others in L2 (plots). Moreover, all models demonstrate improvements when incorporating external character knowledge, as shown in Fig. 5 (right), highlighting the importance of character understanding in

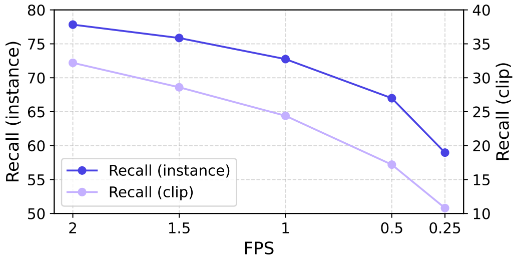


Figure 6: Visual fact recall at varying frame rates. Recall (instance): percentage of recalled visual facts compared to the total visual facts; Recall (clip): percentage of movie clips where all visual facts are recalled.

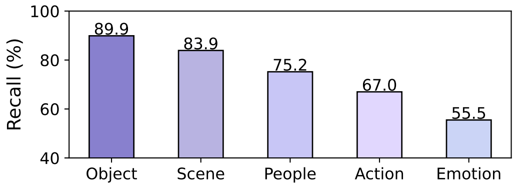


Figure 7: Visual fact recall across various categories.

Table 2: Results of character face recognition test.

Model	# GT	# Predict.	# Correct	Precision
GPT-4V	1,066	1,094	477	43.6
ArcFace	1,066	713	341	47.8

movie narrating. However, despite these gains, all models still exhibit limited task performance, far from being directly applicable, indicating a need for continued research and development. Qualitative results can be found in Appendix D.2.

4.2 Analysis

Given the limited model performance observed in benchmarking experiments on movie narration, we aim to investigate challenges that models encounter and provide insights for future improvements. We analyze these challenges from the perspectives of visual perception and text generation.

4.2.1 Visual Perception

The visual information that a model can perceive from videos depends on (1) its input capacity, i.e., how many frames it can process, and (2) its ability to comprehend the visual inputs. We conduct analytical experiments to pinpoint the essential difficulties related to both aspects, using GPT-4V as a representative for analysis. Given a movie clip $V = \{f_i\}_{i=1}^{L}$ and its ground truth narration y, we first extract the atomic visual facts $A = \{a_j\}_{j=1}^{N}$ mentioned in y using an LLM (GPT-4). We then query GPT-4V to identify the visual elements in A that exist in a particular frame f_i through a binary classification, represented as $R_i = \{a_j\}_{j=1}^{M}, R_i \subseteq A$. Aggregating R_i across all frames gives the total visual elements GPT-4V can recall from V. Our analysis covers 500 clips randomly chosen from the Movie101v2 test set, sampled at 2 FPS. We identify several challenges as follows:

Input Capacity Limitation. From the 2,954 visual facts in the test narrations, GPT-4V recalls 77.8% when processing the full 2 FPS input. However, as the frame rate decreases, especially below 1 FPS, recall drops sharply (Fig. 6), indicating a significant information loss due to limited input capacity. Moreover, a human evaluation of 1,000 movie clips sampled at 1 FPS shows that 23.1% lack the necessary information for accurate narration generation. These results highlight that a limited input capacity represents an early bottleneck for models' visual perception. As most current LVLMs feature a limited visual context length (e.g., 16 frames in VideoChat-2), their ability to perceive long movie clips (e.g., $\geq$ 30 seconds) faces notable challenges. Therefore, improving models to handle longer contexts remains a foremost goal.

Visual Comprehension Limitation. As GPT-4V achieves only a 77.8% recall rate on basic visual recognition tasks, which can further constrain its narration performance, we investigate the specific drawbacks concerning visual comprehension. We divide the visual facts into five categories: objects, scenes, people, actions, and emotions, for a more detailed analysis. As shown in Fig. 7, the model performs better in recognizing objects and scenes but struggles with people-related visual facts, particularly actions and expressions. Since understanding characters and their behaviors is crucial for interpreting movie plots, improving models' comprehension in these aspects deserves further research.

Face Recognition Limitation. Beyond basic visual facts, we also assess the model's ability to identify characters in movie clips based on cast portraits. Following a similar process, given a movie clip and a list of character portraits, we task GPT-4V with identifying the characters appearing in the clip. The ground truth characters are extracted from the reference narration. Table 2 shows the results from 500 randomly chosen clips sampled at 1 FPS. While GPT-4V can generally accurately *detect* characters from the video (matching the ground truth in terms of count), it struggles to *identify* who are the specific characters, achieving only 43.6%

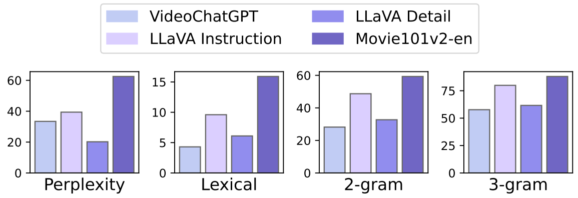


Figure 8: Linguistic properties of different datasets.

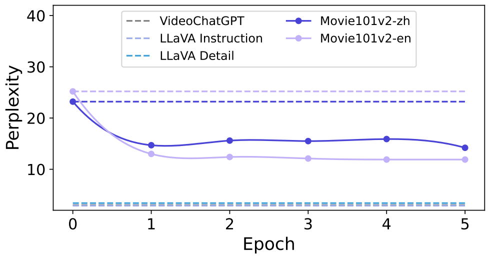


Figure 9: VideoChat-2 perplexity on different datasets. Dotted lines represent the zero-shot perplexity of the model, while solid lines denote the perplexity of the model fine-tuned on Movie101v2.

precision. It also underperforms when compared to ArcFace (Deng et al., 2019), a specialized face recognition model. This limitation in re-identifying cast characters further constrains the model's ability to generate accurate movie narrations.

4.2.2 Textual Generation

In addition to visual perception, we analyze challenges related to generating textual outputs. From both linguistic and model-fitting perspectives, we observe specific challenges for generating narration texts in Movie101v2.

Data Linguistics. We compare the linguistic properties of our movie narration data with those of other datasets commonly used to train LVLMs (Maaz et al., 2023; Liu et al., 2023), using 1,000 samples from each. We evaluate perplexity (a measure of textual fluency and complexity calculated as the exponentiated average negative log-likelihood of a sequence as observed by an evaluator model like GPT-2 (Radford et al., 2019)) and n-gram diversity (the ratio of unique n-grams to the total n-gram counts). As shown in Fig. 8, Movie101v2 has the highest GPT-2 perplexity and greater lexical diversity (1-gram) and n-gram diversity, indicating more complex narration text compared to other datasets. This complexity presents additional learning challenges for models.

Model Fitting. We further examine model fitting challenges using perplexity as a measure of how well the model fits the data. Lower perplexity indicates a better fit. We demonstrate the perplexity of VideoChat-2 on different datasets in Fig. 9. The model starts with the highest zero-shot perplexity on Movie101v2, suggesting unfamiliarity with the narration text. While perplexity decreases after training, it remains much higher than for other datasets, underscoring the specific challenges in achieving a good fit on complex movie narrations.

5 Related Works

Understanding movies by AI has attracted considerable research interest (Zhu et al., 2020; Chen et al., 2023a; Song et al., 2024, 2023; Li et al., 2023c). MovieNet (Huang et al., 2020) provides rich annotations of actors and scenes, supporting various proxy tasks like detection, identification, segmentation, and cinematic style prediction. CMD (Bain et al., 2020) focuses on key scenes coupled with high-level semantic descriptions. YMS (Dogan et al., 2018) and SyMoN (Sun et al., 2022) provide plot summaries from movies and TV series for multimodal story comprehension. For a narrative understanding of movies, M-VAD (Torabi et al., 2015) and MPII-MD (Rohrbach et al., 2015), combined into LSMDC (Rohrbach et al., 2017), provide video-aligned movie narrations. PTVD (Li et al., 2023a) offers detailed human-written plot descriptions. MAD (Soldan et al., 2022) and TVC (Lei et al., 2020), originally designed for tasks such as temporal grounding and text-video retrieval, have also demonstrated their value in narration generation. Together, these datasets enable extensive exploration of movie understanding (Han et al., 2023b,a, 2024; Argaw et al., 2023; Zhang et al., 2023). Our work builds upon the recently proposed Movie101 (Yue et al., 2023), enhancing the data, task definitions, and baseline methods to further support research on generating movie narrations.

6 Conclusion

In this work, we present Movie101v2, a large-scale, bilingual dataset designed to advance the development of automatic movie narration generation, making the following contributions: (1) We create a comprehensive dataset using a scalable, repeatable data collection pipeline. (2) We outline a clear task roadmap for achieving the long-term goal of movie narration, and propose corresponding metrics for task evaluation. (3) We benchmark various state-of-the-art models and investigate essential difficulties to motivate future improvements. Our findings reveal a significant gap between the current model capabilities and the requirements for generating applicable movie narrations, underscoring the importance of further research to enable AI-driven solutions that can enhance the movie-watching experience for visually impaired individuals.

Limitations

This work introduces an enhanced benchmark for automatic movie narration generation. To accommodate current technological constraints, we have simplified the task of movie narration by focusing on individual movie clips. On the one hand, this simplification makes the step-by-step development of deployable movie narration systems more feasible. However, on the other hand, it may constrain more bold and ambitious research aimed at achieving the ultimate goal in a single leap.

Ethics Statement

Movie101v2 is designed to advance research in automatic movie narration generation, with the goal of improving accessibility for visually impaired individuals. We address potential ethical concerns as follows:

Data. Our dataset consists exclusively of Chinese movies, which may introduce a lack of diversity and potential cultural biases. While most movie narration datasets in current literature are based on English-language movies, we believe that Movie101v2 adds valuable representation to the field. Besides, our dataset contains no personally identifiable information or offensive content.

Crowdsourcing. Our data refinement involved a modest amount of crowdsourcing (manual correction of movie cast lists). We compensated annotators with fair wages according to local standards.

Copyrights. The movies in our dataset are publicly available on Xigua Video, and our data collection compiles with the service contract of the website[2]. To respect copyright, we will release our dataset under highly restrictive permissions, limiting its use strictly to academic research purposes.

References

Dawit Mureja Argaw, Joon-Young Lee, Markus Woodson, In So Kweon, and Fabian Caba Heilbron. 2023. Long-range multimodal pretraining for movie understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13392–13403.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*.

Jieting Chen, Junkai Ding, Wenping Chen, and Qin Jin. 2023a. Knowledge enhanced model for live video comment generation. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pages 2267–2272. IEEE.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2023b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms.

DeepSeek-AI. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE.

Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross. 2018. A neural multi-sequence alignment technique (neumatch). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8749–8758.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.

Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. 2023a. Autoad ii: The sequel-who, when, and what in movie audio description. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13645–13655.

[2] https://www.ixigua.com/robots.txt

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2023b. Autoad: Movie description in context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18930–18940.

Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. 2024. Autoad iii: The prequel-back to the pixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18164–18174.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. 2024. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Chen Li, Xutan Peng, Teng Wang, Yixiao Ge, Mengyang Liu, Xuyuan Xu, Yexin Wang, and Ying Shan. 2023a. Ptvd: A large-scale plot-oriented multimodal dataset based on television dramas. *arXiv preprint arXiv:2306.14644*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2023b. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2023c. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *ArXiv preprint*, abs/2304.08485.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. Videogpt+: Integrating image and video encoders for enhanced video understanding.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

OpenAI. 2022. Introducing whisper. `https://openai.com/research/whisper`.

OpenAI. 2023. Introducing chatgpt. `https://openai.com/blog/chatgpt`.

PaddleOCR. 2022. Paddleocr. `https://github.com/PaddlePaddle/PaddleOCR`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3202–3212. IEEE Computer Society.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123:94–120.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. 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*, pages 5026–5035.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan Lu, Jenq-Neng Hwang, et al. 2023. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*.

Zhende Song, Chenchen Wang, Jiamu Sheng, Chi Zhang, Gang Yu, Jiayuan Fan, and Tao Chen. 2024. Moviellm: Enhancing long video understanding with ai-generated movies. *arXiv preprint arXiv:2403.01422*.

Yidan Sun, Qin Chao, Yangfeng Ji, and Boyang Li. 2022. Synopses of movie narratives: a video-language dataset for story understanding. *arXiv preprint arXiv:2203.05711*.

Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *ArXiv preprint*, abs/1503.01070.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zihao Yue, Qi Zhang, Anwen Hu, Liang Zhang, Ziheng Wang, and Qin Jin. 2023. Movie101: A new movie understanding benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4669–4684.

Chaoyi Zhang, Kevin Lin, Zhengyuan Yang, Jianfeng Wang, Linjie Li, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. Mm-narrator: Narrating long-form videos with multimodal in-context learning. *arXiv preprint arXiv:2311.17435*.

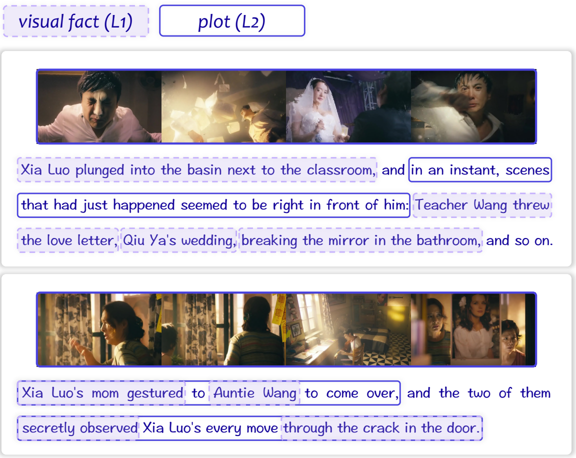
Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yutao Zhu, Ruihua Song, Zhicheng Dou, Jian-Yun Nie, and Jin Zhou. 2020. ScriptWriter: Narrative-guided script generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8647–8657, Online. Association for Computational Linguistics.

Table 3: Left: Percentage of narrations that contain errors; Right: Human evaluation and back-translation (BT, BLEU@4) results of translated narrations. Human rating 1-5: *terrible*, *poor*, *fair*, *good*, *excellent*.

Dataset	Dial.	Text	Name	Avg.	Human	BT
Movie101	0.7	3.3	2.3	2.1	-	-
Movie101v2	0.0	3.7	0.0	1.2	4.56	34.3



Xia Luo plunged into the basin next to the classroom, and in an instant, scenes that had just happened seemed to be right in front of him: Teacher Wang threw the love letter, Qiu Ya's wedding, breaking the mirror in the bathroom, and so on.

Xia Luo's mom gestured to Auntie Wang to come over, and the two of them secretly observed Xia Luo's every move through the crack in the door.

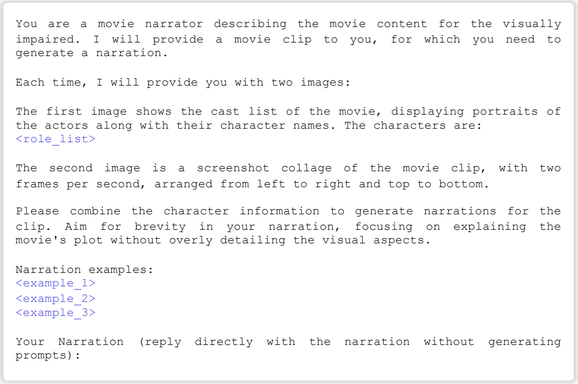
Figure 10: Examples of L1 and L2 narrative elements.

A Data Quality Analysis

With data quality as our top priority, we have meticulously refined our data pipeline through extensive manual checks. We conduct a quality analysis comparing both Movie101 and our newly collected data, using 300 random samples from each. We first manually check the errors in narration texts, including character dialogue remnants, textual mistakes, and character name mismatches. As shown in Table 3, our automatically refined data demonstrates quality comparable to the manually refined Movie101. Additionally, we evaluate the quality of the automatically translated English narrations through human rating (by three annotators) and a back-translation test, where English narrations are translated back into Chinese. Both evaluations, as seen in Table 3, confirm the high quality of the automatic translation.

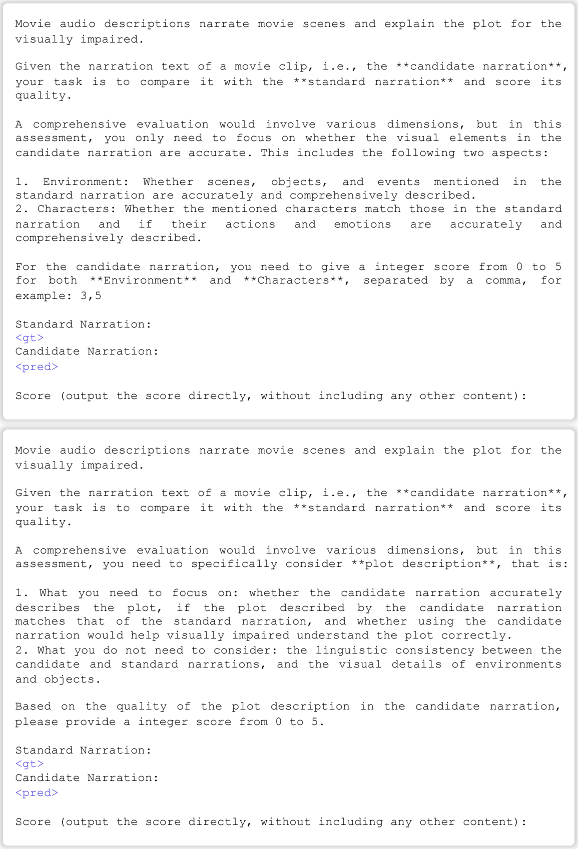
B L1 and L2 Differences

We illustrate the difference between L1 and L2 narrations in Fig. 10. L1 content focuses on basic visual facts that are directly observable in individual frames, while L2 content captures events and plots that are developed by combining L1 elements.



```
You are a movie narrator describing the movie content for the visually
impaired. I will provide a movie clip to you, for which you need to
generate a narration.

Each time, I will provide you with two images:

The first image shows the cast list of the movie, displaying portraits of
the actors along with their character names. The characters are:
<role_list>

The second image is a screenshot collage of the movie clip, with two
frames per second, arranged from left to right and top to bottom.

Please combine the character information to generate narrations for the
clip. Aim for brevity in your narration, focusing on explaining the
movie's plot without overly detailing the visual aspects.

Narration examples:
<example_1>
<example_2>
<example_3>

Your Narration (reply directly with the narration without generating
prompts):
```

Figure 11: Prompts for GPT-4V to generate narrations.



```
Movie audio descriptions narrate movie scenes and explain the plot for the
visually impaired.

Given the narration text of a movie clip, i.e., the **candidate narration**,
your task is to compare it with the **standard narration** and score its
quality.

A comprehensive evaluation would involve various dimensions, but in this
assessment, you only need to focus on whether the visual elements in the
candidate narration are accurate. This includes the following two aspects:

1. Environment: Whether scenes, objects, and events mentioned in the
standard narration are accurately and comprehensively described.
2. Characters: Whether the mentioned characters match those in the standard
narration and if their actions and emotions are accurately and
comprehensively described.

For the candidate narration, you need to give a integer score from 0 to 5
for both **Environment** and **Characters**, separated by a comma, for
example: 3,5

Standard Narration:
<gt>
Candidate Narration:
<pred>

Score (output the score directly, without including any other content):
```

```
Movie audio descriptions narrate movie scenes and explain the plot for the
visually impaired.

Given the narration text of a movie clip, i.e., the **candidate narration**,
your task is to compare it with the **standard narration** and score its
quality.

A comprehensive evaluation would involve various dimensions, but in this
assessment, you need to specifically consider **plot description**, that is:

1. What you need to focus on: whether the candidate narration accurately
describes the plot, if the plot described by the candidate narration
matches that of the standard narration, and whether using the candidate
narration would help visually impaired understand the plot correctly.
2. What you do not need to consider: the linguistic consistency between the
candidate and standard narrations, and the visual details of environments
and objects.

Based on the quality of the plot description in the candidate narration,
please provide a integer score from 0 to 5.

Standard Narration:
<gt>
Candidate Narration:
<pred>

Score (output the score directly, without including any other content):
```

Figure 12: Prompts for producing L1-Score (top) and L2-Score (bottom).

C Implementation Details

C.1 Models

Video Models. Models that process video input directly include VideoGPT+, VideoChat-2, VideoLLaMA 2, Qwen2-VL, and InternVL2. For these models, if not specified, we set a visual context length of 16 frames, with up to 5 main actor portraits prepended to video frames, taking into account input capacity constraints. In the case of VideoGPT+, which processes video frames into 4-frame chunks, we provide up to 4 actor portraits. For Qwen2-VL, which accommodates simultaneous video and image input, we provide actor portraits as separate direct image inputs alongside the video. We train the models on the 46.0K narration paragraphs for 3 epochs and evaluate their performance on the 10 test set movies. All other training configurations are consistent with the official documentation[3][4][5][6][7].

Multi-image Models. Models that process multiple images include Qwen-VL and GPT-4V. For Qwen-VL, we keep $K \leq 4$ segments per video and concatenate the frames sampled at 1 FPS within each segment. Since the model resizes input images to 448px squares, we apply a carefully designed strategy to concatenate both the frames and character portraits while maintaining a balanced aspect ratio. All other training details follow the official configurations for the third training stage of Qwen-VL[8]. GPT-4V handles image understanding through configurable high- and low-resolution settings[9]. In the high-resolution mode, images are scaled to fit within a predetermined size and divided into 512px-square patches, and each patch is represented as a sequence of tokens. For frame images, we apply the high-resolution setting and tailor our frame concatenation strategy to align with the patching approach. We apply the low-resolution setting for portrait images. The API version used is `gpt-4-0314`. We detail the prompt that guides the model to generate movie narrations in Fig. 11.

ArcFace. In Section 4.2, we evaluate the character face recognition performance of ArcFace. The character faces are first detected and aligned using MTCNN (Zhang et al., 2016), where the minimum face size is set to 30 pixels. We then extract their face features using the ArcFace model with a ResNet-50 (He et al., 2016) backbone. Cosine similarities between the face features from frames and portraits are calculated to identify characters detected in the videos.

[3] https://github.com/mbzuai-oryx/VideoGPT-plus

[4] https://github.com/OpenGVLab/Ask-Anything/tree/main/video_chat2

[5] https://github.com/DAMO-NLP-SG/VideoLLaMA2

[6] https://github.com/QwenLM/Qwen2-VL

[7] https://internvl.github.io/blog/2024-07-02-InternVL-2.0/

[8] https://github.com/QwenLM/Qwen-VL

[9] https://platform.openai.com/docs/guides/vision

Table 4: Model performance evaluated by GPT-3.5-turbo.

	Model	w/o character				w/ character			
		L1-Env	L1-Char	**L1**	**L2**	L1-Env	L1-Char	**L1**	**L2**
ZH	InternVL2	**21.6**	34.3	27.9	38.1	23.9	38.7	31.3	**40.0**
	CogVLM2-video	21.3	**37.0**	**29.1**	36.4	22.5	34.5	28.5	37.4
	Qwen-VL	17.3	30.7	24.0	**39.2**	20.7	35.1	27.9	33.6
	Qwen2-VL	21.1	34.8	27.9	35.2	24.2	39.8	32.0	37.1
	GPT-4V	-	-	-	-	**34.9**	**46.5**	**40.7**	36.4
EN	VideoGPT+	14.7	28.6	21.6	**38.0**	18.2	34.1	26.1	**40.3**
	Videochat2	6.8	23.0	14.9	33.3	7.6	24.5	16.0	34.1
	Videollama2	13.8	28.6	21.2	37.2	19.1	34.0	26.5	39.2
	Qwen-VL	8.6	24.7	16.7	33.4	12.5	29.0	20.8	35.7
	Qwen2-VL	**17.3**	**29.1**	**23.2**	37.2	21.7	34.8	28.2	39.2
	GPT-4V	-	-	-	-	**27.7**	**35.0**	**31.4**	35.4

C.2 Evaluation

We introduce two evaluation metrics, L1-Score and L2-Score, in Section 3.2. The prompts guiding the LLMs to produce these scores are provided in Fig. 12. All the LLMs adopt a zero temperature for deterministic decoding.

D Additional Results

D.1 GPT-3.5 Results

In addition to the evaluation results obtained using open-source LLMs in Fig. 5, we also provide the results evaluated using GPT-3.5-turbo (`gpt-3.5-turbo-1106`) in Table 4. These results align closely with those in Fig. 5, indicating the consistency of our evaluation framework across different evaluators. However, one exception is that in the English setting, GPT-3.5 shows a clear preference for the output of GPT-4V, which contrasts with Llama-3.1's evaluation.

D.2 Qualitative Results

We present qualitative results from the baseline models in Fig. 13. The models can often successfully recognize static visual elements, such as scenes, objects, and characters within movie frames. However, they struggle with accurately describing fine-grained events or visual details, such as human interactions or subtle emotions of characters, which often require advanced visual perception and the ability to capture visual changes across consecutive frames. The gap between model-generated narrations and human-created ones underscores the need for further advancements to achieve practical automatic movie narration.

Sis Liu, I'd better be going.
"西虹市一大杰出青年"授予仪式
Award Ceremony of the Only Outstanding Youth of Xihong City

GT: 牧良逢害羞地说着，柳烟倒是在离别时刻轻轻拥抱着他，牧良逢神态复杂，一旁喝茶的国军看着热闹，像是担心被传闲话，柳烟又退开。

InternVL2: 柳烟紧紧地抱住他，泪水在眼眶里打转。她亲吻着他的脸颊和额头，依依不舍地说道。

CogVLM2-Video: 柳烟一把抱住他。

Qwen-VL: 柳烟看着牧良逢，眼神中充满了爱意，她伸出手摸了摸牧良逢的头发，随后又将手伸向了牧良逢的胸口，牧良逢低头看着柳烟，随后缓缓抬起头，

Qwen2-VL: 牧良逢紧紧地抱住她，两人依偎在一起久久不愿分开，在一旁看着他们的士兵们也纷纷投来羡慕的目光。

GPT-4V: 在一个古朴的村庄，牧良逢和柳烟之间发生了一段深情的对话。牧良逢身穿军装，看起来严肃而坚定，而柳烟则身着传统服饰，面容柔和。两人的目光交汇，似乎在无声地传递着彼此的情感。柳烟轻轻地抚摸着牧良逢的手，她的动作充满了关切和不舍。随后，牧良逢紧紧握住柳烟的手，他们的眼神中流露出深深的依恋和即将分离的忧伤。在这一刻，周围的世界仿佛静止了，只剩下他们两个人的情感交流。最终，牧良逢缓缓松开手，转身离开，而柳烟目送着他的背影，脸上的表情复杂，既有坚强也有无奈。

GT: 豪宅主人刘轩站在屋内，看着离自己越来越近的富豪们，向前走了几步，笑着张开双臂，向前来做客的大家表示欢迎。

InternVL2: 李若兰和手下们站在门口，看着西装革履的二人

CogVLM2-Video: 刘轩穿着一件亮晶晶的浴袍走进大堂，他笑容满面地迎接大家

Qwen-VL: 刘轩站在门口，看见李若兰和八哥走进酒店，他急忙跑上前去，一把拉住李若兰的手。

Qwen2-VL: 刘轩穿着一身金色西装走进酒店大堂，他面带微笑地向大家打招呼。

GPT-4V: 在一座华丽的酒店大堂，刘轩身着一件闪亮的西装，显得自信满满地走下宽阔的楼梯。他的步伐稳健，目光坚定，似乎正赴一个重要的约会。随着他的接近，大堂中的人群开始注意到他的到来，有人开始窃窃私语。刘轩走到大堂中央，停下脚步，四下环视。他的表情从容不迫，但眼神中透露出一丝寻找的意图。突然，他的目光定格在一个方向，脸上露出了微笑。接着，镜头转向八哥和珊珊，他们正站在大堂的一角，看起来似乎在等待着刘轩的到来。八哥身穿一件花哨的衬衫，旁边的珊珊则穿着一件时尚的外套，两人交谈着，气氛轻松。最后，刘轩走向八哥和珊珊，三人会合，似乎是老朋友重逢。他们的对话充满了笑声和友好的招呼，显然是在开心地讨论着即将发生的事情。

***GT:** The scene transitioned from the exterior of West Rainbow City to the indoor West Rainbow City Annual Influential Figure Awards Ceremony, where Wang Duoyu received the trophy from the guests.*

VideoGPT+: The award ceremony ended, and Wang Duoyu was surrounded by reporters. Xia Zhu stood aside, watching everything. The director took the award from the host and handed it to Wang Duoyu, who took it and looked at it with a smile.",

VideoChat-2: The scene shifts to a stage where a man is handing a trophy to another man.

VideoLLaMA 2: The scene shifts to the Xiangyang City, Hubei Province, where the only outstanding youth in the city is being awarded. Wang Duoyu is dressed in a suit, holding a plaque and a certificate, with a proud smile on his face."

Qwen-VL: The host announced the winner of the "Outstanding Young Person of Xinhong City" award, and Wang Duoyu stood up to receive the award

Qwen2-VL: The audience cheered as Wang Duoyu walked onto the stage to receive his award. He was wearing a suit with a tie, holding an award plaque in both hands. The plaque had Chinese characters on it: “Xinhong City‘s Outstanding Youth Award”

GPT-4V: *Wang Duoyu stands on a stage, addressing an attentive audience with a microphone in hand. The scene shifts to Xia Zhu, who looks on with a mixture of concern and curiosity. Zhuang Qiang, in a moment of realization, exchanges a significant glance with Mr. Jin, who is seated and holding a tablet displaying a message. The message is then passed to Liu Jiannan, who takes a careful look at the tablet. The atmosphere is tense, suggesting a pivotal moment of revelation or decision-making among the characters*

***GT:** JC swung around the head of the vicious dog and threw it to the ground, then turned and ran away before it could catch*

***VideoGPT+:** JC was knocked down by a fierce dog, and the other two dogs rushed over. He quickly picked up the bag and ran, but the dogs caught up*

***VideoChat-2:** The dog bites the man's arm and pulls him down.*

***VideoLLaMA 2:** The two dogs were about to pounce on JC, but were stopped by Simon's command. JC was scared and ran away, while the two dogs chased after him.*

***Qwen-VL:** He leaped over the bushes, and the two dogs chased after him. He ran through the grass, dodging the bushes, and then turned to face the two dog*

***Qwen2-VL:** The police officer was thrown to the ground by JC, who then chased after him with his dog. The police officer struggled to get up but was quickly caught again by JC.*

***GPT-4V:** JC is suddenly tackled to the ground by an assailant. Struggling amidst the dense foliage, he manages to throw his attacker off. A large dog then charges towards him, but JC skillfully dodges, using the maze-like hedges to his advantage. The dog pursues relentlessly as JC navigates the labyrinth of greenery, looking for an escape. He stumbles but quickly recovers, continuing his desperate sprint as the dog gains on him. With a burst of speed, JC leaps over a low hedge, narrowly evading the snapping jaws of his canine pursuer*

Figure 13: Qualitative results from our baseline models on Movie101v2.