MVSBench: A Benchmark for Multi-modal Video Comprehension with Enriched Context

Anonymous ACL submission

Abstract

In recent years, notable progress has been made in Multi-modal Large Language Models (MLLMs), along with the development of various benchmarks assessing their comprehension abilities. However, most benchmarks focus on visual information understanding and QA tasks, lacking the ability to evaluate performance in complex scenarios that involve audio information and other additional context. To address this gap, we introduce the ***M**ulti-modal **V**ideo **S**tory generation **Bench**mark*, referred to as ***MVSBench***, a benchmark designed to evaluate MLLMs' ability to generate narrative-style captions for long videos enriched with auxiliary information. We propose an automatic dataset construction pipeline that reduces manual annotation while ensuring fairness and reliability through filtering techniques and state-of-the-art models. Experiments indicate that current state-of-art MLLMs perform poorly under our evaluation metrics, highlighting significant limitations in generating narratives enriched with auxiliary information. To address these challenges, we propose a novel framework, ***M**ovie-to-**S**tory (M2S)*, which outperforms other MLLMs by over 13% on MVSBench.

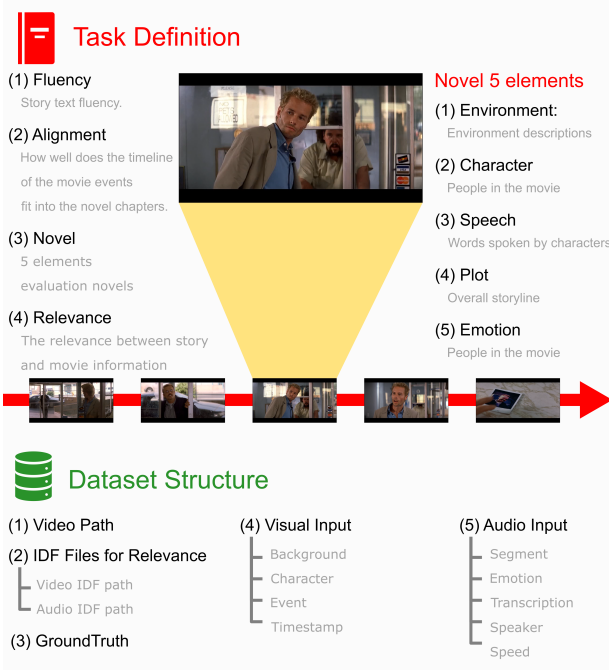


Figure 1: Overview of MVSBench and Dataset Structure: MVSBench covers four subdomains and includes 11 tasks, (Novel has 5 tasks shown in figure and relevance has 4 tasks, see detail definitions in section 3) providing a comprehensive evaluation framework for multi-modal video understanding with enriched context.

1 Introduction

In recent years, many MLLMs (Alayrac et al., 2022; Zhu et al., 2023; Li et al., 2024a; Huang et al., 2023; Li et al., 2020; OpenAI, 2022) have effectively used video (Touvron et al., 2023; Devlin et al., 2019; Dosovitskiy et al., 2021) and audio encoders (Radford et al., 2022; Chen et al., 2022b,a) to extract multimodal information and generate text. However, as MLLMs continue to advance, a critical challenge emerges: how can we effectively evaluate their comprehension and text generation capabilities? This challenge is particularly relevant in assistive media applications, such as improving accessibility for individuals who are deaf or blind. Existing subtitles primarily convey basic visual information but lack comprehensive auditory descriptions, limiting the viewing experience for this audience. Generating enriched video descriptions that integrate both visual and audio elements can address this gap. But assessing their effectiveness is difficult without a reliable benchmark. Developing effective evaluation methods is essential to advance MLLMs and verify their ability to produce meaningful multimodal narratives.

Existing benchmarks for MLLMs primarily adopt a question-answering (QA) format (Yu et al., 2024; Xu et al., 2023; Xie et al., 2024; Fu et al., 2024a; Xu et al., 2016; Liu et al., 2024; Cheng

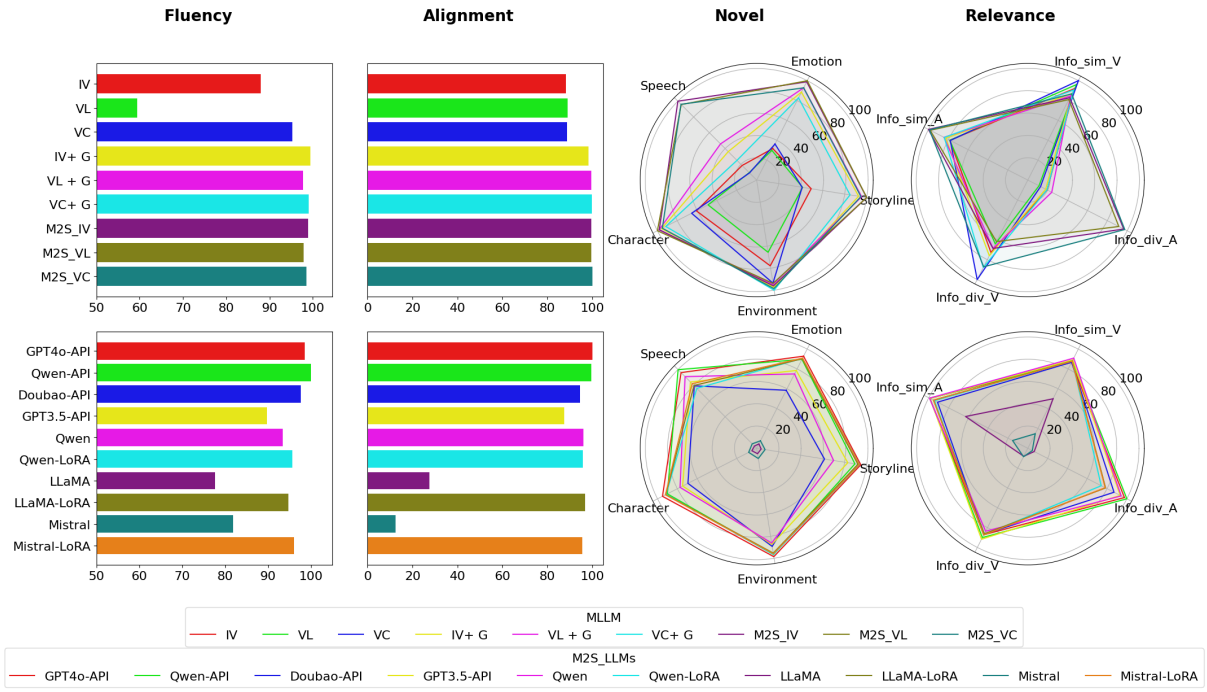


Figure 2: Performance overview of LLMs and MLLMs on MVSBench: Each column represents an evaluation domain. Each row shows the chosen model performance on defined task. For baseline MLLMs, VideoChat2 performs best. For open-source LLMs, GPT-4o achieves the highest overall performance. For closed-source LLMs, Qwen performing best. VC means Videochat2. IV means Internvideo2. VL means VideoLLava2. _ means different baseline MLLM.

et al., 2024), focusing on static image understanding. While benchmarks such as MVBench (Li et al., 2024b) extend evaluation to temporal tasks, they remain inadequate for assessing long-video comprehension and fail to incorporate rich auxiliary information such as audio. Although some studies, such as AV-SUPERB (Tseng et al., 2024), attempt to integrate audio information with visual information, they primarily focus on the evaluation of audio. Moreover, current benchmarks typically produce objective, template-like outputs, lacking the stylistic complexity of narrative storytelling. Furthermore, many benchmarks rely heavily on manual annotations, which are resource-intensive and time-consuming.

To overcome these limitations, we introduce a novel benchmark, **M**ulti-modal **V**ideo **S**tory generation **Bench**mark (**MVSBench**), which emphasizes the integration of diverse auxiliary information (e.g., audio features) to generate long-video descriptions in a narrative and information-rich style.

Our approach introduces an innovative automatic data generation pipeline to enhance existing video-text datasets (Li et al., 2020) by incorporating detailed audio and visual information. For instance, original video captions such as "a car is driving" are expanded into richer narratives like "Tom, dressed in a black suit, sings: 'Oh, beautiful sun ...' while Jessica drives the car...".

For audio processing, we extract attributes derived from the audio source, including Automatic Speech Recognition (ASR) (Kheddar et al., 2024) outputs, emotions, and sound events, using state-of-the-art audio models such as Whisper (OpenAI, 2022). Similarly, for visual descriptions, we use advanced vision models (Bai et al., 2023; Chen et al., 2024) to generate detailed frame-level annotations while ensuring temporal consistency.

Our benchmark shows two key advantages. First, it significantly reduces reliance on manual annotation by leveraging automated processes and open-source tools. Second, it incorporates enriched information that is often implicit in videos. For instance, generating a high-quality, story-like video summary requires detailed descriptions of the environment, character names, speech, and actions. Our benchmark provides a comprehensive evaluation of narrative text quality in open-world, long-duration scenarios.

As shown in Fig. 2, we evaluate several state-of-the-art MLLMs (Li et al., 2024a; Lin et al., 2024; Wang et al., 2022, 2024b) on the MVSBench bench-

mark, identifying significant performance gaps. For instance, models like VideoChat2 (Li et al., 2024b) struggle to integrate audio features effectively into high-quality narrative outputs and perform poorly on long-video tasks.

To address these limitations, we introduce **M**ovie-to-**S**tory (**M2S**), a novel framework composed of MLLM and LLM, specifically designed to align with our benchmark's requirements and improve text generation quality. Experimental results show that our baseline outperforms existing MLLMs across our evaluation metrics.

All models, datasets, and evaluation frameworks are publicly available to facilitate future research and advancements in the field.

2 Related Works

2.1 MLLM Introduction

The evolution of Large Language Models (LLMs) has accelerated research into MLLMs (Lyu et al., 2023; Lee et al., 2024; Fu et al., 2024b), aiming to integrate diverse modalities like text, vision, and audio. Early models, such as Flamingo (Alayrac et al., 2022) and PaLM-E (Driess et al., 2023), present strong performance in multi-modal tasks by combining visual and textual modalities. Subsequent open-sourced efforts, including LLaVA (Liu et al., 2023) and MiniGPT-4 (Zhu et al., 2023), have expanded the scope of multi-modal instruction tuning, while VideoChat (Li et al., 2024a) and VideoChatGPT (Maaz et al., 2024) extended these ideas to dynamic video tasks by utilizing ChatGPT-generated annotations.

Building on these works, recent studies have introduced advanced approaches to address the tasks of video understanding. MMAD (Ye et al., 2024b) integrates video, audio events, and text information to generate concise and informative descriptions. Similarly, the Distilling Vision-Language Models on Millions of Videos framework (Zhao et al., 2024) adapts vision-language models for video-language tasks, enabling the generation of high-quality captions while enhancing semantic and contextual understanding.

The Video-LLaMA2 (Cheng et al., 2024) model aligns audio, video, and textual data into a unified space using the pre-trained ImageBind (Girdhar et al., 2023) module. Instead of training an audio-text dataset, it utilizes a video-text encoder to indirectly convert audio into text.

Video storytelling (Li et al., 2020) emphasizes the creation of text summaries for events by selecting important frames using a reinforcement learning-based Narrator model. This approach incorporates contextual embeddings through a Residual Bidirectional RNN (ResBRNN) resulting in more detailed and coherent descriptions.

Our work builds on these advancements by addressing key limitations of existing MLLMs, particularly in processing long-duration videos, integrating auxiliary information, and generating stylistically rich captions. By incorporating innovative dataset enhancement techniques and robust evaluation metrics, our framework establishes a new benchmark for multi-modal understanding and narrative-driven text generation.

2.2 Benchmark Introduction

Traditional Vision-Language (VL) benchmarks (Goyal et al., 2017; Kay et al., 2017; Xu et al., 2016, 2017; Xiao et al., 2021) have primarily focused on QA-style evaluations, emphasizing tasks such as multi-modal retrieval and vision-based question answering. More recent benchmarks have expanded this scope to assess broader multi-modal capabilities.

For instance, OwlEval (Ye et al., 2024a) and SEED-Bench (Li et al., 2023) introduce evaluation metrics that emphasize comprehensive multi-modal reasoning. In the video domain, benchmarks such as Perception Test (Pătrăucean et al., 2023) evaluate multi-modal video perception and reasoning, while FunQA (Xie et al., 2024) evaluates models with humorous and counterintuitive content to improve performance in video-based reasoning.

MVBench (Li et al., 2024b) stands out by defining over 20 tasks to evaluate MLLMs' performance on diverse scenarios, especially temporal reasoning. The research on Synchronized Video Storytelling (Yang et al., 2024) presents a novel methodology that incorporates supplementary advertising keywords to enhance the generation and evaluation of storytelling and advertising content, offering major insights for the discipline. These advancements inform our framework, which aims to address gaps in evaluating narrative-driven, multi-modal tasks.

3 MVSBench

In this section, we explore the details of our MVSBench in depth. Initially, we formulate the multi - modal story generation tasks, as graphically presented in Figure 1. Subsequently, we automate the

generation of video-caption pairs for evaluation, as detailed in the following sections. And overview of the pipeline is shown in Figure 3.

3.1 Task Definition

In the MVSBench framework, we use a text-to-text approach to create story tasks, transforming formatted texts into coherent narratives, such as novels. While most MLLM benchmarks focus on converting video or audio into brief captions, they evaluate vision and audio separately. In contrast, MVSBench integrates both modalities, enabling a unified evaluation of multi-modal storytelling with enriched context.

We first outline core tasks related to video and audio understanding based on previous benchmarks, then expand them into detailed narratives. This leads to the development of story tasks requiring a full understanding of both video and audio. We outline 4 subdomains with 11 specific tasks below:

Fluency. Evaluates the overall fluency of the description. **Alignment.** Verifies whether the narrative structure follows the reference event order. **Novel.** (1) *Environment:* Assesses the quality of environment descriptions. (2) *Character:* Focuses on character descriptions, including clothing and facial expressions. (3) *Speech:* Checks for the presence and quality of speech descriptions. (4) *Storyline:* Evaluate the development, consistency, and logical coherence of the narrative. (5) *Emotion:* Assesses the quality of emotional descriptions. **Relevance.** (1) *Visual Similarity*: Measures the degree to which key visual knowledge is retained in the generated text. (2) *Audio Similarity*: Measures the degree to which key audio knowledge is retained in the generated text. (3) *Visual Diversity*: Assesses the degree to which essential visual information from the knowledge base is used. (4) *Audio Diversity*: Assesses the degree to which essential audio information from the knowledge base is used.

3.2 Automatic Dataset Generation

Based on the definition of the Enriched Video-Captions Generation task, we collect popular existing datasets and annotate the videos. Specifically, we introduce a novel automatic dataset generation process (Fig. 6), which efficiently converts open-source videos into a structured format for both evaluation and fine-tuning.

Datasets Collection. We select several high-quality existing datasets. **(1) VideoInstruct100K** (Muhammad Maaz and Khan, 2023) is a high-quality video conversation dataset that incorporates semi-automatic techniques to assist with annotation. **(2) Video_Story** (Gella et al., 2018) is a new large-scale dataset designed to advance multi-sentence video description, presenting a novel challenge in this domain. **(3) MSR_VTT** (Xu et al., 2016) is a large-scale benchmark for video understanding. It includes 10K web video clips (38.7 hours) with 200K clip-sentence pairs.

Data Pre-processing. we utilize the following tools to convert original videos into a structured format: **(1) Video splitting**: we segment all long videos into 20-second clips, which is typically sufficient to generate narrative text of around 1,000 words. **(2) FFmpeg** (FFmpeg Developers, 2023): We use it to extract keyframes, capturing essential visual information for later processing.

Data Generation. Some datasets provide only brief descriptions, often lacking depth. Additionally, original captions do not include enriched information such as audio content. Thus, expanding reference captions is crucial for evaluating this task. We use state-of-the-art MLLMs and LLMs to extract detailed visual and audio information. For example, Qwen_VL2 (Wang et al., 2024a) captures frame-level visual details, providing a structured reference for event timeline development. For visual information, we employ MLLMs like VideoChat2 (Li et al., 2024b) to extract environment settings, character descriptions, and story development summaries. For audio information, models such as FunASR (et al., 2023a) and Whisper (OpenAI, 2022) capture audio events, emotions, speaker features, and ASR transcriptions. Finally, after obtaining these key components, we use ChatGPT-4o (et al., 2024b) to integrate the background information and generate coherent story-like video summaries. And NLTK (Bird and Loper, 2004) and SpaCy (Honnibal and Montani, 2017) are used in text processing. We construct our proposed dataset in this way from Video_Story and VideoInstruct100K, named **A-VidStory**.

3.3 Prompt Design

To ensure story coherence, we design a detailed prompt for generating as shown in the Fig. 6. This prompt guides LLMs to carefully analyze video and audio content, emphasizing timestamps and other key factors to maintain narrative flow. For the downstream LLM, the prompt encourages full utilization of original visual and audio information to enhance accuracy. It also helps structure the

storyline, ensuring that the final captions are fluent and align with the provided narrative development. By using this guidance framework, we aim to enhance the quality of generated stories, making them more aligned with real-world scenarios and user expectations.

And we also design an effective prompt for evaluation as shown in Fig. 7.

3.4 Automatic Evaluation Metrics

Traditional NLP metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), fail to capture deeper narrative elements, often overlooking audio cues, character details, and other contexts. To address this, we define a knowledge base method that integrates additional visual and audio information.

Supervised Storyline Score(SS). This score measures the fidelity of the generated story against a predefined canonical events order. It assesses how closely the produced narrative aligns with key event orders, character behaviors, and thematic elements detailed in the original outline.

Intra-story Repetition (ISR). ISR evaluates the coherence repetition rate of generated texts. Inspired by related work (Yang et al., 2024), we define a keyword-triplet method to assess fluency at three levels: within a sentence, between different sentences, and overall caption. Excessive triplet repetition indicates unnatural phrasing and redundancy, negatively affecting fluency.

Information Similarity (InfoSim). (Yang et al., 2024) It measures the alignment between the knowledge points in the story and the knowledge repository. A higher similarity score indicates that the generated story effectively incorporates relevant knowledge.

Information Diversity (InfoDiv). (Yang et al., 2024) It evaluates the breadth of the knowledge used in the story. A higher diversity score indicates that the story incorporates a wide range of knowledge points, avoiding over-reliance on a small subset of information.

Qualitative Metrics. This method uses GPT-4 (et al., 2024a) and carefully designed prompts to effectively assess caption quality. It ensures a consistent and scalable evaluation of generated captions. The prompt for GPT-4 is shown in Fig. 7. The definded metrics are: **(1) Unsupervised Storyline Score (USS)** evaluates how reasonable, deep, and engaging the story's progression is and whether the story feels well-developed and logical. **(2) Environment Score (EvS)** assesses how well the text sets the scene, considering environmental richness, sensory details, and atmospheric quality. **(3) Emotion Score (EmS)** evaluates the depth and authenticity of emotional expression, focusing on how effectively the text conveys characters' emotions and overall tone. **(4) Speech Score (SpS)** measures the realism and vividness of human speech, assessing whether the dialogue is engaging and natural. **(5) Character Score (CaS)** examines the depth and consistency of character portrayals, focusing on appearance, personality, and behavior.

4 M2S pipeline

After establishing MVSBench, we first evaluated the MLLMs listed in Table. 1 on our benchmark. However, results show that current MLLMs struggle to generate detailed long-form text, often lacking audio information. To address this gap, we propose a robust MLLM-LLM pipeline, Movie-to-Story (**M2S**).

4.1 Synthesized Instruction-Tuning Dataset

Current MLLMs struggle to generate story-like captions with enriched information. To bridge this gap, a downstream LLM can be utilized to process enriched inputs and generate the final captions. However, as shown in Table 1, some LLMs, such as LLaMA3-3B, perform poorly without fine-tuning, likely due to limited diversity in instruction-tuning data. To address this, we synthesize a fine-tuning dataset, comprising 10K samples from MSR_VTT, incorporating both visual and audio information to enhance fine-tuning effectiveness.

The process is similar to A-VidStory construction process. But we employ ChatGPT-4o to generate additional audio descriptions for some videos without audio. And to ensure coherence between vision and audio information, we condition ChatGPT-4o on the extracted vision features as contextual background.

The instruction-tuning dataset and A-VidStory are converted to a uniform format shown in Fig. 5. There are three key components: `{video}`, `{visual_input}`, and `{audio_input}`. The first key stores the video file path. The second key contains structured visual information, including environment, storyline, characters, and timestamps. The third key represents audio features such as emotion, transcription, speed, and speaker identity, aligned with time segments.

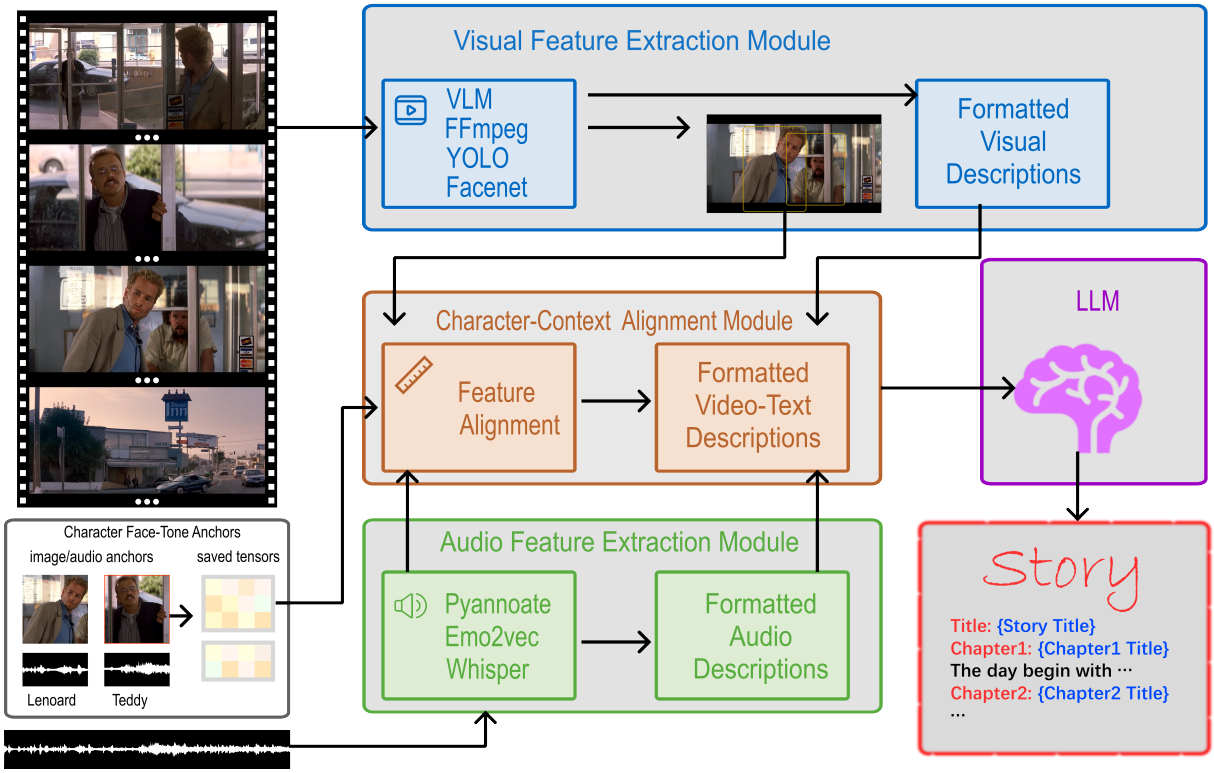


Figure 3: Overview of M2S pipeline: M2S consists of multiple modules designed to extract enriched multimodal information for generating detailed captions.

4.2 Visual Feature Extraction Module

This module extracts key visual information from videos. We use current state-of-the-art MLLMs like VideoChat2 (Li et al., 2024b) as the baseline model. With a structured prompt, MLLMs identifies essential elements, including environment settings, character descriptions, and story events summaries.

To process long videos, we segment them into 5-second clips. Visual features are extracted from each clip and aggregated into a structured format for subsequent LLM input. Additionally, YOLOv8 (Yaseen, 2024) is employed for person segmentation, while FaceNet (Schroff et al., 2015) extracts facial features to assist in character name alignment and description mapping.

4.3 Audio Feature Extraction Module

This module employs several open-source models to extract audio features. Pyannote (et al., 2019) segments the audio into speech clips and extract acoustic features for speaker alignment.

For transcription, we use Whisper (OpenAI, 2022). Emotion analysis is conducted using Emotion2Vec (et al., 2023b) from FunASR (et al., 2023a). Additionally, we compute word speed to assess speaking rate. The final output includes timestamps, transcriptions, emotion labels, speech rate, and speaker references, forming a comprehensive audio representation.

4.4 Character-Context Alignment Module

Identifying a person's name from their visual appearance in a video is challenging. Existing methods, such as MMAD (Ye et al., 2024b), have shown strong performance in person re-identification using image-based features. However, aligning a speaker's audio content with their visual representation requires a multimodal approach.

To address this, we integrate both visual and audio features. Preprocessed anchor data containing a person's name, facial features, and speech attributes are used to establish a strong mapping between the visual and audio domains. Additionally, vision and audio features are linked with their respective textual descriptions.

This ensures that names are correctly assigned to the corresponding individuals in both visual and spoken content, enhancing downstream text generation.

Model	avg	*Sim_V*	*Sim_A*	*Sim_avg*	*Div_V*	*Div_A*	*Div_avg*	EmS	SpS	CaS	EvS	USS	Overall	ISR(↓)	SS(↓)	Storylength
InternVideo2	48.20	**64.6**	31.1	47.9	30.7	15.0	22.8	20.9	12.0	35.2	66.9	24.0	31.8	15.1	46.4	619
InternVideo2+GPT4o	55.72	61.0	32.7	46.8	**32.1**	13.9	23.0	58.3	23.4	55.5	**83.6**	44.9	53.2	**4.0**	40.4	794
M2S(InternVideo2, GPT4o)	**64.04**	58.2	**39.6**	**48.9**	29.0	**74.5**	**51.8**	**65.1**	**65.0**	**56.8**	82.6	**48.4**	**63.6**	4.6	**39.5**	762
VideoLLava2	40.96	**67.5**	31.1	**49.3**	27.1	8.7	17.9	19.7	6.1	28.4	56.4	20.2	26.2	42.7	45.9	763
VideoLLava2+GPT4o	56.08	59.5	33.4	46.5	**29.0**	18.6	23.8	61.2	29.8	57.0	**81.3**	46.2	55.2	**5.6**	**39.5**	855
M2S(VideoLLava2, GPT4o)	**62.84**	56.9	**38.9**	47.9	26.1	**70.7**	**48.4**	**65.9**	**62.2**	**57.9**	80.6	**48.4**	**63.0**	**5.6**	**39.5**	806
VideoChat2	51.22	**70.3**	30.9	**50.6**	**42.2**	9.9	26.1	23.9	5.9	38.0	80.4	20.0	33.6	8.1	46.1	646
VideoChat2+G	56.16	63.7	33.3	48.5	37.1	14.9	26.0	54.5	16.2	53.1	**86.0**	41.1	50.2	**4.5**	39.4	840
M2S(VideoChat2, GPT4o)	**64.74**	60.9	**39.2**	50.0	36.8	**75.3**	**56.1**	**61.1**	**62.4**	**55.1**	84.7	**45.9**	**61.8**	4.9	**39.3**	850
GPT3.5-API*	59.78	**62.7**	**39.1**	**50.9**	**38.7**	72.0	55.3	51.6	54.7	43.3	75.1	40.1	53.0	13.4	46.9	1110
Doubao-API*	59.88	61.0	36.1	48.6	36.6	67.2	51.9	38.5	51.5	40.2	76.7	29.9	47.4	5.8	42.7	664
Qwen-API*	**65.12**	62.0	**39.1**	50.5	38.0	77.2	**57.6**	**59.1**	**64.9**	**52.2**	**83.1**	43.3	**60.5**	**3.5**	**39.5**	753
LLaMA3-LoRA*	61.16	61.9	37.7	49.8	36.4	60.6	48.5	59.5	51.7	53.1	81.9	44.6	58.2	8.5	42.2	876
Mistral-LoRA*	**61.56**	61.7	37.7	49.7	**36.9**	60.3	48.6	**59.9**	53.3	**53.3**	82.4	45.2	**58.8**	**7.3**	42.0	855
Qwen-LoRA*	61.06	61.9	37.3	49.6	**36.9**	57.2	47.0	59.7	49.4	53.0	**83.0**	**45.3**	58.1	7.6	41.8	870
LLaMA3*	25.68	35.1	24.9	30.0	3.4	5.0	4.2	2.9	1.8	2.8	4.3	2.0	2.7	25.1	83.4	504
Mistral*	20.62	10.3	6.1	8.2	3.6	3.5	3.5	4.9	3.7	4.7	8.0	3.7	5.0	21.1	92.5	308
Qwen*	61.26	**63.9**	**39.3**	**51.6**	35.3	**72.8**	**54.0**	49.4	**59.1**	44.8	74.1	33.9	52.2	9.8	**41.7**	650

Table 1: Experiment Results: For example M2S(InvideoVideo2, GPT4o) represents one case of our M2S framework, where InvideoVideo2 is used as the MLLM and GPT4o serves as the LLM. The results emphasize the limitations of current MLLMs in generating enriched video captions in complex scenarios. LoRA enables us to fine-tune local LLMs to achieve performance comparable to closed-source models. The symbol $*$ denotes the LLM chosen as the downstream model of M2S. All models follow the strategy of segmented clips.

5 Experiment

5.1 Experiment Settings

Implementation Details. To measure the performace of leading LLM sand MLLMs on MVS-Bench, we design a detail experiment workflow and test extensively state-of-art models. **(1) Baseline LLM**: For the closed-source LLMs, we employ tht API of GPT-4o, GPT-3.5 (Brown et al., 2020), Doubao, Qwen. For the open-source LLM, we test LLaMA-3.2-3B, Qwen2.5-7B-Instruct, Mistral-7B-v0.1. Detailed information is shown in Appendix A. **(2) Baseline MLLM**: We test InternVideo2 (Wang et al., 2024b), Video-LLaVA2 (Lin et al., 2024) and Videochat2 (Li et al., 2024b) on our benchmark.

We finetune the open-source LLM on 10K instruction-tuning dataset,to evaluate its performance. We use 90% of the data for fine-tuning and the remaining 10% as the test set. The proposed framework is trained for 5 epochs with a learning rate of $5e^{-5}$. The LoRA (et al., 2021) parameters are set to $r = 32$ and $\alpha = 16$. Training the closed-source LLMs took approximately 15 hours on three 3090 GPUs.

5.2 Results Analysis

Evaluation results on MVSBench, shown in Table 1, indicate that current MLLMs struggle with story-like caption generation. M2S outperforms the base MLLM combined with GPT-4o by over 24% and surpasses the base MLLM alone by over 29% in the diversity metric. In qualitative metrics, our pipeline achieves a 25% improvement compared to the base MLLM. M2S improves overall average scores by over 13% compared to the base MLLMs. These improvements presents M2S's ability to generate richer narratives by integrating both visual and audio information. In the similarity metrics, these models exhibit comparable performance, demonstrating that our pipeline effectively retains key information.

Baseline MLLMs often produce shorter outputs with higher repetition rates and struggle with maintaining logical coherence in long-video contexts. In contrast, M2S generates longer, more structured, and fluent captions while preserving key details.

For further details and qualitative comparisons, please refer to Figure 4.

5.3 LoRA Result Analysis

LoRA significantly enhances the instruction-following ability of local models, making them comparable to API-based models like GPT-4o. Before LoRA, LLaMA and Mistral performed poorly, often failing to generate meaningful outputs, while Qwen already demonstrated strong instruction adherence. After fine-tuning, all models showed improvement. LLaMA3-3B presents a 50% gain of qualitative score after finetuning. More details are provided in Table 1.

In qualitative evaluations, LoRA-tuned models surpassed GPT-3.5 and Doubao by over 5%, demonstrating the effectiveness of our pipeline. Compared to end-to-end MLLM training, our approach is both feasible and resource-efficient. By using MLLMs for text extraction and LLMs for novel generation, we enable the scalable production of long multimodal narratives while avoiding excessive computational overhead.

5.4 Ablations Study

Table 1 presents the ablation study of our pipeline. Compared to base MLLMs combined with GPT-4o, M2S achieves a 7.8% increase in five GPT scores, approximately a 1.4% boost in the audio similarity score, and a 24% improvement in the audio diversity score. In the speech score, our pipeline outperforms other models by over 32%. These results indicate that incorporating the audio module effectively enhances overall performance. This highlights the essential role of the audio module in enhancing overall performance. We also evaluated different closed-source LLM APIs. GPT-4o achieved the best results, particularly in the five core metrics, while GPT-3.5 and Doubao performed poorly in overall scores.

6 Conclusion

Our paper introduces MVSBench, a comprehensive benchmark designed to evaluate MLLM's multimodal story generation capabilities. We also propose a pipeline, M2S, that performs better than the leading models on the MVSBench benchmark. Our extensive analysis provides valuable insights into the design of MLLMs for multimodal story generation, especially in scenarios rich in additional information. Despite these advances, there are still some limitations to our current approach. We aim to address these issues in future work to enhance the assessment framework and further improve MLLM's performance in complex, information-rich environments.

Limitations

One limitation of our approach is the lack of an end-to-end framework that directly processes video input and produces enriched textual descriptions. Instead, we rely on separate components for multimodal extraction and text generation. For storyline matching, our metric struggles with complex scenarios, such as evaluating the relevance of storylines spanning multiple time segments and modeling intricate narrative structures. Additionally, our evaluation primarily compares generated texts without fully incorporating visual frame feature, which could further refine accuracy assessment. Since our benchmark emphasizes the role of audio information, it inherently requires videos to contain relevant auditory elements. Addressing these limitations is an important direction for future work.

Internvideo:

The video shows a man in a black suit and a man in a beige suit talking to each other through a glass window. The man in the black suit is holding a cell phone and the man in the beige suit is holding a bag. They are both standing in front of a sign that says \"Desert Inn.\" A car is passing by in the background.

VideoLLaVA:

The video shows a man walking down a street and talking to another man. The man in the suit is seen walking away from the camera and then back again. The man in the suit is also seen walking down the street and talking to another man.

VideoChat:

"The video starts with a blurry image of a man in a suit standing in front of a glass door. The scene then shifts to a clearer view of the same man, who is now inside the building, looking out through the glass door. He is wearing a light-colored suit and a blue shirt. The man appears to be in a conversation with another person who is not visible in the frame. The background shows a parking lot with cars and a building with a sign that reads \"NO PARKING ALLOWED.\" The man's facial expressions change from neutral to smiling as he continues to look out the door. The scene changes again to show the man from a different angle, this time with a blurred background, indicating motion or a change in focus.",

M2S:

Chapter 1: A Fateful Encounter
Lennord stood by the glass door of the building, his light-colored suit and blue shirt giving him an air of quiet professionalism. The sign on the door, "NO PERSONS ALLOWED," caught the eye of anyone passing by. Through the glass, the blurry outline of a parking lot shimmered in the midday sun, cars idling in the distance. Inside, the atmosphere was tense, despite the mundane surroundings of the office. Lennord's posture was rigid, his eyes focused as if waiting for something—or someone.

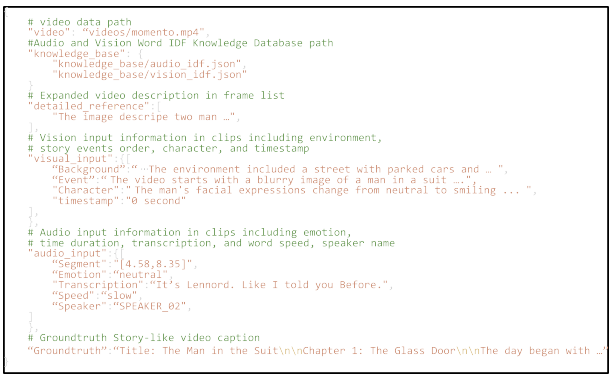
Suddenly, the door opened, revealing Teddy, a man in blue clothes, his glasses catching the light. His face lit up with a wide smile, a flash of happiness overtaking him as he called out in a slow, deliberate voice, "Lenny!"

Teddy's joy was clear, but beneath the surface, there was a subtle undertone of something more complex—perhaps a tinge of frustration or even anger, though barely noticeable.

Lennord's sharp blue eyes narrowed slightly. He turned towards Teddy, his voice calm, measured. "It's Leonard. Like I told you before." His tone was slow and neutral, though there was an unmistakable firmness in his words.

...

Figure 4: Comparison of Generation Results: Our pipeline achieves better performance by incorporating enriched contextual information.



```
# video data path
"video": "videos/momento.mp4",
#Audio and Vision Word IDF Knowledge Database path
"knowledge_base":[
    "knowledge_base/audio_idf.json",
    "knowledge_base/vision_idf.json"
],
# Expanded video description in frame list
"detailed_reference":[
    "The image descripe two man …",
],
# Vision input information in clips including environment,
# story events order, character, and timestamp
"visual_input":[
    "Background":"…the environment included a street with parked cars and … ",
    "Event":" The video starts with a blurry image of a man in a suit …",
    "Character":" The man's facial expressions change from neutral to smiling ... ",
    "Timestamp":"0 second"
],
# Audio input information in clips including emotion,
# time duration, transcription, and word speed, speaker name
"audio_input":[
    "Segment":"[4.58,8.35]",
    "Emotion":"neutral",
    "Transcription":"It's Lennord. Like I told you Before.",
    "Speed":"slow",
    "Speaker":"SPEAKER_02",
],
# Groundtruth Story-like video caption
"Groundtruth":"Title: The Man in the Suit\n\nChapter 1: The Glass Door\n\nThe day began with …"
```

Figure 5: Example of dataset structure.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *Preprint*, arXiv:2308.12966.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022a. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. *Preprint*, arXiv:2202.00874.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. 2022b. Beats: Audio pre-training with acoustic tokenizers. *Preprint*, arXiv:2212.09058.

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. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *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. *Preprint*, arXiv:2406.07476.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model. *Preprint*, arXiv:2303.03378.

Edward J. Hu et al. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hervé Bredin et al. 2019. pyannote.audio: neural building blocks for speaker diarization. *Preprint*, arXiv:1911.01255.

OpenAI et al. 2024a. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI et al. 2024b. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Zhifu Gao et al. 2023a. Funasr: A fundamental end-to-end speech recognition toolkit. *Preprint*, arXiv:2305.11013.

Ziyang Ma et al. 2023b. emotion2vec: Self-supervised pre-training for speech emotion representation. *Preprint*, arXiv:2312.15185.

FFmpeg Developers. 2023. FFmpeg tool (Version 4.4.1). Software available from `http://ffmpeg.org/`. Accessed: 2023-10-05.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Shaoqi Dong, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Ran He, Rongrong Ji, Yunsheng Wu, Caifeng Shan, and Xing Sun. 2024b. Vita: Towards open-source interactive omni multimodal llm. *Preprint*, arXiv:2408.05211.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*,

pages 968–974, Brussels, Belgium. Association for Computational Linguistics.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all. *Preprint*, arXiv:2305.05665.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "something something" video database for learning and evaluating visual common sense. *Preprint*, arXiv:1706.04261.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. *Preprint*, arXiv:2302.14045.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset. *Preprint*, arXiv:1705.06950.

Hamza Kheddar, Mustapha Hemis, and Yassine Himeur. 2024. Automatic speech recognition using advanced deep learning approaches: A survey. *Information Fusion*, 109:102422.

Seon-Ho Lee, Jue Wang, David Fan, Zhikang Zhang, Linda Liu, Xiang Hao, Vimal Bhat, and Xinyu Li. 2024. Nowyousee me: Context-aware automatic audio description. *Preprint*, arXiv:2412.10002.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *Preprint*, arXiv:2307.16125.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. 2020. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024a. Videochat: Chat-centric video understanding. *Preprint*, arXiv:2305.06355.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *Preprint*, arXiv:2306.09093.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

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

OpenAI. 2022. Whisper: A general-purpose speech recognition model. `https://cdn.openai.com/papers/whisper.pdf`. Accessed: 2023-10-06.

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. Association for Computational Linguistics.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Preprint*, arXiv:2305.13786.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 815–823. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yuan Tseng, Layne Berry, Yi-Ting Chen, I-Hsiang Chiu, Hsuan-Hao Lin, Max Liu, Puyuan Peng, Yi-Jen Shih, Hung-Yu Wang, Haibin Wu, Po-Yao Huang, Chun-Mao Lai, Shang-Wen Li, David Harwath, Yu Tsao, Shinji Watanabe, Abdelrahman Mohamed, Chi-Luen Feng, and Hung yi Lee. 2024. Av-superb: A multi-task evaluation benchmark for audio-visual representation models. *Preprint*, arXiv:2309.10787.

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. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. *Preprint*, arXiv:2403.15377.

Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. 2022. Internvideo: General video foundation models via generative and discriminative learning. *Preprint*, arXiv:2212.03191.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions. *Preprint*, arXiv:2105.08276.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. Funqa: Towards surprising video comprehension. *Preprint*, arXiv:2306.14899.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5288–5296.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *Preprint*, arXiv:2306.09265.

Dingyi Yang, Chunru Zhan, Ziheng Wang, Biao Wang, Tiezheng Ge, Bo Zheng, and Qin Jin. 2024. Synchronized video storytelling: Generating video narrations with structured storyline. *Preprint*, arXiv:2405.14040.

Muhammad Yaseen. 2024. What is yolov8: An in-depth exploration of the internal features of the next-generation object detector. *Preprint*, arXiv:2408.15857.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024a. mplug-owl: Modularization empowers large language models with multimodality. *Preprint*, arXiv:2304.14178.

Xiaojun Ye, Junhao Chen, Xiang Li, Haidong Xin, Chao Li, Sheng Zhou, and Jiajun Bu. 2024b. MMAD:multi-modal movie audio description. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11415–11428, Torino, Italia. ELRA and ICCL.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. *Preprint*, arXiv:2308.02490.

Yue Zhao, Long Zhao, Xingyi Zhou, Jialin Wu, Chun-Te Chu, Hui Miao, Florian Schroff, Hartwig Adam, Ting Liu, Boqing Gong, Philipp Krähenbühl, and Liangzhe Yuan. 2024. Distilling vision-language models on millions of videos. *Preprint*, arXiv:2401.06129.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

A Formula

A.1 Information Scores

$$\text{InfoSim} = \frac{1}{2|s_i|}\sum_{s_i}\left(\max_{k\in K} f_k^T f_{s_i} + \frac{1}{|W(s_i)|}\sum_{w\in W(s_i)}\max_{k\in K} f_k^T f_w\right) \quad (1)$$

$$\text{InfoDiverse} = \frac{1}{|K|}\left|\bigcup_{s_i}\{k_t \in K \mid \max_{w\in W(s_i)\cup S_i} f_{k_t}^T f_w > 0.9\}\right| \quad (2)$$

where $W(s_i)$ represents all words in sentence s_i. f_{s_i}, f_k, and f_w refer to the normalized embeddings of sentence s_i, knowledge point k, and segmented word w, respectively.

$$R_\lambda = \frac{\sum_{s_i,s_j\in S_\lambda}\sum_{t\in T_\lambda(s_i,s_j)} w_\lambda \cdot f_\lambda(t,s_i,s_j)}{\sum_{s_i,s_j\in S_\lambda}|T_\lambda(s_i,s_j)|} \quad (3)$$

where:

- λ denotes the level of granularity for repetition rate calculation:
  - $\lambda =$ overall (global repetition rate)
  - $\lambda =$ inter (sentence-to-sentence repetition rate)
  - $\lambda =$ intra (intra-sentence repetition rate)
- S_λ represents the scope of the calculation:
  - $\lambda =$ overall: all trigrams in the text.
  - $\lambda =$ inter: all sentence pairs (s_i, s_j).
  - $\lambda =$ intra: a single sentence s_i.
- $T_\lambda(s_i, s_j)$ represents the set of trigrams:
  - $\lambda =$ overall: all trigrams T.
  - $\lambda =$ inter: shared trigrams between two sentences, $T_{s_i} \cap T_{s_j}$.
  - $\lambda =$ intra: trigrams within a single sentence T_{s_i}.
- w_λ is a weighting factor:
  - $\lambda =$ overall: $w_\lambda = 1$.
  - $\lambda =$ inter: $w_\lambda = 2$ (to normalize the pairwise count).
  - $\lambda =$ intra: $w_\lambda = 1$.
- $f_\lambda(t, s_i, s_j)$ counts the occurrence of trigram t:
  - $\lambda =$ overall: $f(t) - 1$ (total occurrences minus unique count).
  - $\lambda =$ inter: $\min(f(t, s_i), f(t, s_j))$ (minimum count between two sentences).
  - $\lambda =$ intra: $f(t, s_i) - 1$ (internal repetition in a single sentence).

A.2 Alignment Scores

SS is the score for the storyline, which quantifies the mismatch rate between events in the video and the story.

$$\text{SS} = \frac{\text{InversionCount}(f(g,r))}{\text{MaxInversionCount(g)}}$$

Variable Description:

(1) InversionCount : The reverse number of matching index lists (the number of reverse pairs)

(2) g,r: generated text and reference texts.

(3) $f : T_g \times T_r \to R^n$: Cosine Similarity Best Match Index List R^n of texts generated text with respect to reference texts.

- T_g and T_r are text spaces for generated text and reference texts.

- n is number of the sentence/chapter segments in generated text g.

$$\text{f}(g_j, r) = \arg\max_i (\cos(E(r_i), E(g_j)))$$

- $E(\cdot)$: Sentence Transformer Encoder.
- $\cos(\cdot,\cdot)$: Cosine similarity calculation.

(4) MaxInversionCount(g) $= \frac{n(n-1)}{2}$

B Experiments and Analysis

B.1 Novel Analysis

Our evaluation of the novel includes five elements: Emotion, Speech, Character, Environment, Storyline. For a detailed introduction, please refer to the section 3.

According to Tab.5 GPT4o API is the best in Emotion (EmS), Character (CaS), Environment (EvS), and Storyline (USS) while while the best performance on Speech(SpS) is on Qwen-API , indicating that GPT4o and Qwen-API have outstanding abilities in integrating rich information and generating process text.

In Tab.6, our M2S pipeline is better than MLLM-LLM without audio information in all four aspects except for the Environment. This may be because the addition of audio information has squeezed

out some visual information, which is currently included in our environment. If audio descriptions of the environment are added in the future, this part could also be better.

B.2 Fluency and Storyline Analysis

ISR and SS are quantitative metrics used to evaluate the fluency of novel language and the consistency of the overall storyline order. Both metrics indicate that the smaller the better.

According to Tab.3 and Tab.2. (1)ISR: The model without LLM processing will perform worse in terms of language fluency in novels, and the LLaMA score before LoRA will perform worse. This is because the ISR metric evaluates the language repetition within the text and overall. The more repetition, the more inclined the model is to generate the same text, and the unprocessed or untrained model is indeed prone to generating the same text repeatedly, which is in line with our expectations. The special feature is that Videollava's unprocessed text has a high repetition rate. After our observation, we found that the same sentence was repeated many times, which disappeared after LLM processing, and the score was almost the same as the other two MLLMs, reflecting the robustness of our pipeline. (2)SS: The relative difference in SS among all models is not significant, except for LLaMA and Mistral before processing LoRA. This is because these two models had poor instruction obedience before LoRA, making them more prone to descriptions with disordered order or empty strings. The data processed by LLM with good command obedience is very good, indicating that the storyline matches the order of events in the video. Compared to not using LLM processing, directly attaching VLM results in chronological order is not a good method. This may be because VLM results are often shorter, and the probability of similar text appearing in different positions is higher, which affects the evaluation of the order of the storyline.

B.3 Information Metric Analysis

In this section, we can analyze the degree to which LLM preserves information and the importance of audio information intervention. In order to avoid randomness in the experiment, the model was strictly programmed to write novels based on the input information, with all LLM temperatures set to 0.

According to Tab.4, (1) Similarity score: After adding Audio, the sim score for Video information decreases, while the div score for Audio information increases. Qwen perform the best in preserving information similarity. It may tend to generate similar texts with input. While GPT4o and others may tend to use richer expressions such as synonym replacement or sentence re modification. (2) Diversity score: After adding Audio, the div score for Video information decreases slightly, while the div score for Audio information increases significantly. This indicates that after adding audio information to our input, LLM ensures that sufficient audio information is retained during processing. However, due to the fact that the length of the generated text does not change too much see Tab.1, the information in the video section is compressed to a small extent, which affects the diversity of information.

In summary, based on the information metrics infosim and infodiv, we can understand that LLM does retain key information from the video in the generated novels, which is crucial for producing novels that meet the requirements.

B.4 NLP Analysis

As shown in Tab.3 and, we use the generation captions from M2S pipeline with VideoChat2 and GPT4o as the components. Because the original caption is too short to be a good reference. The improvement in NLP metrics demonstrates that LoRA enables local LLMs to achieve performance comparable to closed-source LLMs.

C Figures and Tables

Model	ISR(↓)	SS(↓)
InternVideo2	15.1	46.4
VideoLLava2	42.7	**45.9**
VideoChat2	**8.1**	46.1
InternVideo2+GPT4o	**4.0**	40.4
VideoLLava2+GPT4o	5.6	39.5
VideoChat2+GPT4o	4.5	**39.4**
InternVideo2+GPT4o+Audio module	**4.6**	39.5
VideoLLava2+GPT4o+Audio module	5.6	39.5
VideoChat2+GPT4o+Audio module	4.9	**39.3**

Table 2: Quantitative performance of various base MLLMs with GPT4o as LLM on M2S

Model	ISR(↓)	SS(↓)	Rouge-1(↑)	Rouge-2(↑)	Rouge_L(↑)	BLEU-4(↑)
GPT4o-API	4.9	**39.3**				
GPT3.5-API	13.4	46.9				
Doubao-API	5.8	42.7				
QWen-API	**3.5**	39.5				
QWen	9.8	**41.7**	57.5	29.4	31.6	14.5
Mistral	21.1	92.5	7.1	2.4	3.8	1.0
LLaMA	25.1	83.4	9.7	2.2	6.5	0.6
QWen-LoRA	7.6	41.8	67.5	34.9	34.9	20.4
Mistral-LoRA	**7.3**	42.0	**68.3**	**35.7**	**35.5**	**21.3**
LLaMA-LoRA	8.5	42.2	66.6	33.9	34.2	19.6

Table 3: Quantitative performance of various downstream LLMs on M2S

Model	*Sim_V*	*Sim_A*	*Sim_avg*	*Div_V*	*Div_A*	*Div_avg*
VideoChat2+GPT4o+Audio Module	60.9	39.2	50.0	36.8	**75.3**	**56.1**
VideoChat2+GPT4o	63.7	33.3	48.5	37.1	14.9	26.0
VideoLLava2+GPT4o+Audio Module	56.9	38.9	47.9	26.1	70.7	48.4
VideoLLava2+GPT4o	59.5	33.4	46.5	29.0	18.6	23.8
InternVideo2+GPT4o+Audio Module	58.2	39.6	48.9	29.0	74.5	51.8
InternVideo2+GPT4o	61.0	32.7	46.8	32.1	13.9	23.0
LLaMA-LoRA	61.9	37.7	49.8	36.4	60.6	48.5
Mistral-LoRA	61.7	37.7	49.7	36.9	60.3	48.6
Qwen-LoRA	61.9	37.3	49.6	36.9	57.2	47.0
LLaMA	35.1	24.9	30.0	3.4	5.0	4.2
Mistral	10.3	6.1	8.2	3.6	3.5	3.5
Qwen	**63.9**	**39.3**	**51.6**	35.3	72.8	54.0
GPT3.5-API	62.7	39.1	50.9	**38.7**	72.0	55.3
Doubao-API	61.0	36.1	48.6	36.6	67.2	51.9
Qwen-API	62.0	39.1	50.5	38.0	77.2	57.6
InternVideo2	64.6	31.1	47.9	30.7	15.0	22.8
VideoLLava2	67.5	31.1	49.3	27.1	8.7	17.9
VideoChat2	70.3	30.9	50.6	42.2	9.9	26.1

Table 4: Performance of Open Source LLMs and Baseline (VideoChat) with Audio on Information Metrics

Model	EmS	SpS	CaS	EvS	USS	Overall
GPT4o-API	**3.057**	3.122	**2.753**	**4.236**	**2.294**	**3.092**
GPT3.5API	2.582	2.734	2.164	3.755	2.007	2.648
Doubao-API	1.926	2.575	2.009	3.835	1.494	2.368
QWen-API	2.953	**3.243**	2.610	4.155	2.165	3.025
QWen	2.468	2.955	2.240	3.703	1.694	2.612
Mistral	0.247	0.185	0.234	0.400	0.185	0.250
LLaMA	0.144	0.092	0.140	0.213	0.099	0.137
QWen-LoRA	2.986	2.471	2.648	4.151	2.263	2.904
Mistral-LoRA	2.993	2.663	2.666	4.120	2.259	2.940
LLaMA-LoRA	2.974	2.587	2.657	4.097	2.231	2.909

Table 5: Performance of various LLMs on Qualitative metrics

Model	EmS	SpS	CaS	EvS	USS	Overall
InternVideo2	1.046	0.601	1.760	3.346	1.199	1.590
VideoLLava2	0.987	0.307	1.418	2.818	1.010	1.308
VideoChat2	1.195	0.296	1.898	4.021	1.001	1.682
InternVideo2+GPT4o	2.916	1.172	2.777	4.181	2.243	2.658
VideoLLava2+GPT4o	3.060	1.490	2.848	4.067	2.310	2.755
VideoChat2+GPT4o	2.725	0.812	2.653	**4.299**	2.055	2.509
InternVideo2+GPT4o+Audio Module	3.257	**3.248**	2.842	4.128	2.420	**3.179**
VideoLLava2+GPT4o+Audio Module	**3.294**	3.110	**2.897**	4.032	**2.421**	3.151
VideoChat2+GPT4o+Audio Module	3.057	3.122	2.753	4.236	2.294	3.092

Table 6: Performance of various models with GPT4o as LLM on Qualitative metrics

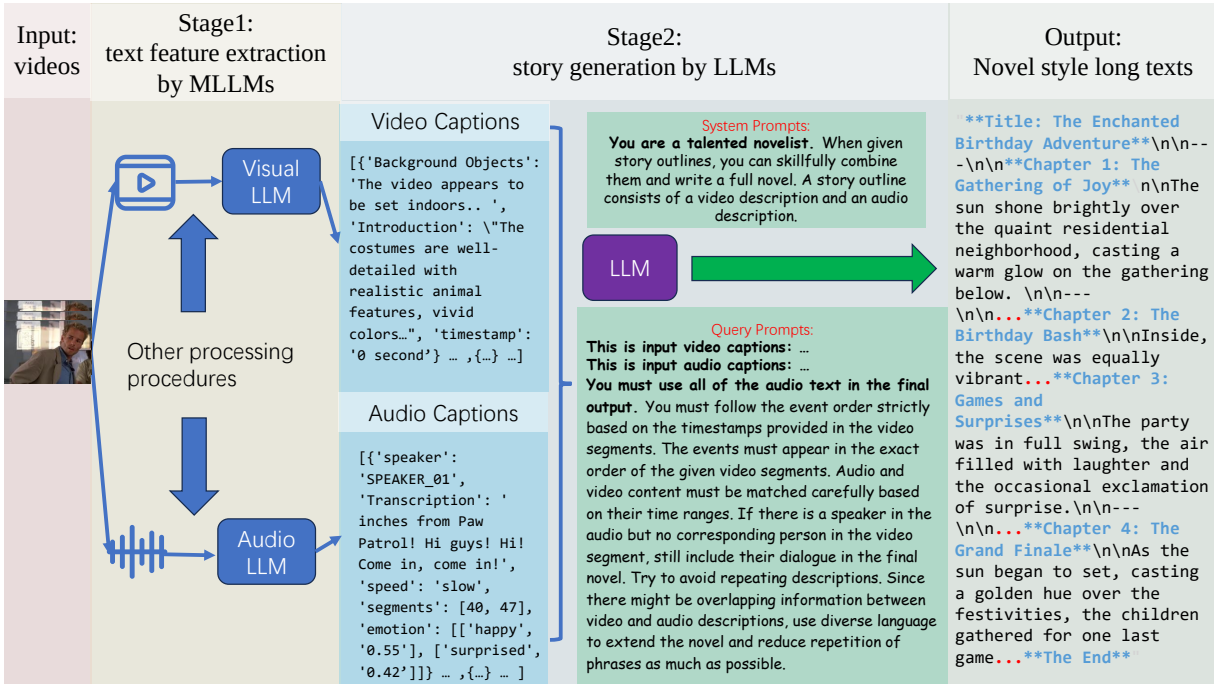


Figure 6: pipeline example with prompts

```
PROMPT = {

    "system_prompt": """
        You are an expert reviewer with advanced knowledge in storytelling and writing
evaluation.
        Your task is to evaluate novel excerpts based on specific storytelling elements
provided by the user.

        Always adhere to the following guidelines:
        - Focus on the user's instructions and evaluate only the requested elements.
        - For "Speech Description," prioritize whether the excerpt includes human speech,
dialogue, or direct quotes, and assess the realism, engagement, and clarity of such
language.
        - Output results in strict JSON format, including integer scores (0-5) for each
category and an overall score.
        - Ensure the JSON format is valid, follows the structure provided, and contains no
extraneous information.

        If there are ambiguities in the user prompt, infer the most reasonable evaluation
criteria based on common storytelling practices. Always explain each score succinctly if
required by the user.
        """,


    "user_prompt": """
            Please evaluate the following novel excerpt based on the five key elements of
storytelling:
                1. **Psychological/Emotional Description**: Does the excerpt convey the
emotions and inner thoughts of characters? How detailed and rich is the emotional
portrayal?
                2. **Speech Description**: Does the excerpt include vivid and expressive
use of human speech or dialogue? Are there realistic and engaging conversations or quotes?
Is the language style engaging and evocative when depicting spoken words?
                3. **Character Description**: Are the characters described in detail? How
well does the excerpt develop or portray the characters?
                4. **Environment Description**: Is the setting described effectively? How
vivid and detailed is the description of the surroundings?
                5. **Plot Development**: How reasonable, deep, and engaging is the
progression of the story's plot? Does the plot feel well-developed and logical?

            For each category, provide a **score as an integer from 0 to 5** (e.g., 0 =
completely absent, 5 = excellent).

            **Output the evaluation strictly in the following JSON format:**
            {
                "psychological_emotional": <integer from 0 to 5>,
                "speech": <integer from 0 to 5>,
                "character": <integer from 0 to 5>,
                "environment": <integer from 0 to 5>,
                "plot": <integer from 0 to 5>,

            }

            Now, please evaluate the following text strictly and professionally:
            """
}
```

Figure 7: GPT prompts for qualitative evaluation