Learning Video Context as Interleaved Multimodal Sequences

Kevin Qinghong Lin[1], Pengchuan Zhang[2], Difei Gao[1], Xide Xia[2], Joya Chen[1], Ziteng Gao[1], Jinheng Xie[1], Xuhong Xiao[3], Mike Zheng Shou[1]✉

[1]Show Lab, National University of Singapore [2]Meta AI [3]DSO National Laboratories

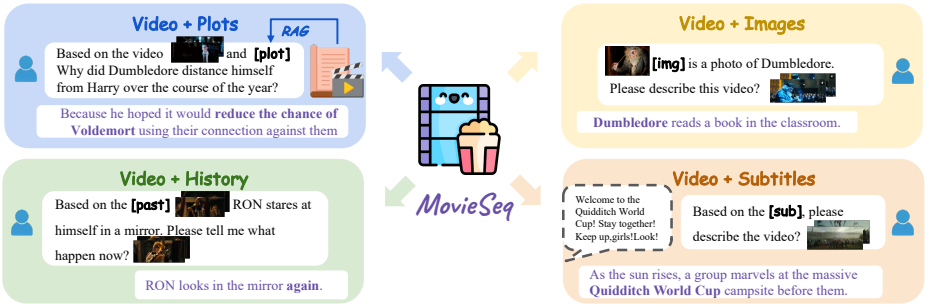


Fig. 1: `MovieSeq` aims to address diverse challenges in understanding video contexts, enabling flexible interleaved multimodal instructions, such as Video+Images (for character identification), Video+Subtitles (for dialogues understanding), Video+Plots (for external knowledge via RAG), and Video+History (for event dependency).

Abstract. Narrative videos, such as movies, pose significant challenges in video understanding due to their rich contexts (characters, dialogues, storylines) and diverse demands (identify who [23], relationship [84], and reason [72]). In this paper, we introduce `MovieSeq`, a multimodal language model developed to address the wide range of challenges in understanding video contexts. Our core idea is to represent videos as interleaved multimodal sequences (including images, plots, videos, and subtitles), either by linking external knowledge databases or using offline models (such as whisper for subtitles). Through instruction-tuning, this approach empowers the language model to interact with videos using interleaved multimodal instructions. For example, instead of solely relying on video as input, we jointly provide character photos alongside their names and dialogues, allowing the model to associate these elements and generate more comprehensive responses. To demonstrate its effectiveness, we validate `MovieSeq`'s performance on six datasets (LVU, MAD, Movienet, CMD, TVC, MovieQA) across five settings (video classifcation, audio description, video-text retrieval, video captioning, and video question-answering). The code will be public at https://github.com/showlab/MovieSeq.

Keywords: Video Understanding · Large Language Models

✉: Corresponding Author.

1 Introduction

Narrative videos, like movie clips, TV series, etc., offer a lens into our diverse world, depicting human stories across history and culture through extended visual streams. These videos pose unique challenges such as character identification [23, 28, 66, 86] (e.g., recognizing 'Who'), situation understanding [7, 72, 77, 84] (e.g., relationship and dialogue), and event dependency [24, 53, 67] (e.g., cause-and-effect over time). Yet, most top-performing video perception models [18, 73, 82] typically prioritize on concise videos, emphasizing atomic objects [1, 61, 65] and actions [21, 31, 33].

Significant progress in understanding video contexts has been made via varied setups such as video recognition [13, 84], video-text retrieval [7, 66], and video question-answering [53, 72]. These advances include: efficient transformer architecture to support extended video durations, applied in tasks like classification [11, 83, 85] and video question-answering [20, 87]; enhanced temporal event association for video-text retrieval [46, 47, 54], localization [25, 47] and video-captioning [14, 34]; audio modality integration for enhanced situational understanding [48, 50, 90]; and the use of external knowledge, as exemplified by [23] in training a character recognition module by actors's databases [28] for audio descriptions. Despite these advancements, prior methods remain task-specific, with design limitations persisting primarily due to: (i) The diversity of contexts (e.g., images, videos, texts, subtitles) and tasks (captioning, retrieval, question-answering, etc). (ii) The tailored technical designs required to adapt to particular settings such as [23] train extra modules to access external knowledge. Having witnessed the advancements in Large Language Models (LLM) [58, 59, 74, 75], the question arises: can we develop a general solution that handles these diverse contexts and needs in videos?

Although LLMs exhibit versatility in natural language processing [58, 74, 75] and multi-modal scenario [5, 49, 94], developing a LLM for complex video understanding is not straightforward. Pioneering efforts such as [38, 39, 49, 94] project single visual input (e.g., image or video) to textual tokens space in conjunction with user queries, as shown in Fig. 2a. Meanwhile, studies like [2,37,78] (see Fig. 2b) introduce interleaved in-context learning to improve the model's few-shot capabilities by providing structured demonstration. However, when applied to narrative videos, which encompass informative contexts , these models with a pre-defined visual-textual template still exhibit limitations due to inflexibility. This highlights the need for a more flexible approach to handling multimodal contexts within videos.

Motivated by these observations, we develop `MovieSeq`, a multimodal language model that is designed for narrative video understanding. Acknowledging the diverse contexts and tasks in videos, our core concept is to embed the videos as interleaved multi-modal sequences. As depicted in Fig. 1 and Fig. 2c, our approach unifies various multimodal contexts (images, subtitles, plots, video history) and tasks into a user-friendly sequence, subsequently processed by the language model. Moreover, one of the key challenges is the lack of instruction-following data for complex videos. We present a packaged solution on how to convert existing video

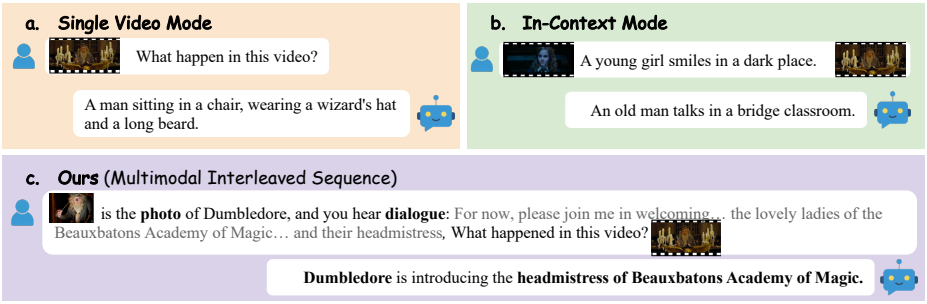


Fig. 2: Comparison between different video-language input modes. (a) Single video input, e.g., [39, 49, 94]. (b) In-context input, e.g., [2, 37], showcasing examples for structured few-shot learning. (c) Our approach, utilizes flexible contexts (e.g., external character images, dialogues, etc) to associate them to produce a comprehensive response.

datasets for interleaved multimodal instruction-following. Finally, we present a decoder-only multi-modal language model, which is trained on our constructed data, to support a variety of task types.

To demonstrate the effectiveness of `MovieSeq`, we conducted experiments across six video benchmarks (LVU [84], MAD [24], Movienet [28], CMD [7], TVC [35], MovieQA [72]) across various settings (video classification, audio description, character identification, video-text retrieval, video captioning, and video question-answering). Additionally, `MovieSeq` facilitates a novel application, allowing users to interact with video using free-form interleaved multi-modal instructions. Overall, our contributions are three folds:

1. We propose `MovieSeq`, a video-language model that embeds videos into interleaved multimodal sequences to flexibly adapt to diverse contexts.
2. We present a package of solutions for converting videos into various forms of interleaved multimodal instructions (e.g., video-images, video-plots, etc).
3. We demonstrate the flexibility and effectiveness of `MovieSeq`, one generative model, which performs well across five tasks and six datasets.

2 Related Work

Movie Understanding seeks to build an interpretation of narrative video streams, which goes beyond basic object and action recognition. This presents challenges in achieving high-level video understanding and reasoning. In the visual unimodal domain, various works [29, 30, 81] have focused on improving network architecture for efficient long-form video recognition. When comes to the multimodal domain, there has been a range of benchmarks focusing on advanced aspects of video contexts, encompassing topics like characters [66, 67], relationships [77, 84], emotions [69], dialogues [7, 35], storylines [28, 88]. To address these challenges, substantial efforts have been made. Studies like [4, 9, 70] primarily enhance video-text alignment along the temporal axis. [34] employs a memory module to improve video captioning's coherency. [48] introduces sound integration to assist video retrieval. Recent AutoAD-II [23] trains a character recognition module using external image databases [28] for precise automatic AD generation.

Notwithstanding this progress, most methods predominantly focus on individual tasks with tailored architectures, illustrating the complexity of video contexts and tasks. Our work seeks one model solution to support various modalities contexts, unifying diverse tasks in a generative manner.

Large Multimodal Models. The advent of LLMs [12, 58, 64] has significantly impacted natural language processing and inspired numerous studies [19, 71, 89] in multimodal domains. Several studies [17, 38, 39, 49, 94] have pioneered Large Multimodal Models by projecting the single visual input (e.g., an image or video) into textual embedded spaces, subsequently aligning them with the LLMs (e.g., Llama [74, 75]) by visual instruction tuning. Additionally, various efforts such as [2, 37, 78], have investigated interleaved vision-text input modes, aiming to augment language models with in-context, few-shot demonstrations. More recently, several studies [15, 92] have focused on enhancing visual spatial understanding e.g., models are capable of producing responses with coordinates. A number of work [42, 68] have proposed efficient architectural solutions to address the memory challenges posed by long video durations.

Nevertheless, the complexity of narrative videos, which include diverse elements such as characters photos, dialogue, and external metadata, continues to pose challenges. Recent work by [44, 91] has proposed the use of GPT-4V [51] to transform video streams into textual document format. In our work, we intend to utilize an open-source language model, e.g., Llama2 [75] with interleaved multimodal instruction tuning for various video applications.

3 Challenges in Video Contexts

In this section, we illustrate challenges and needs associated with understanding contexts in narrative videos. We choose movies as a representative example due to their descriptive storylines, informative metadata such as characters, and extended durations.

(i) Situational Dialogue. While visuals serve as a basic medium for conveying information, achieving a comprehensive understanding of the context demands that the model accommodates audio inputs such as dialogues. This requires the models to associate and interpret the significance of dialogue in conjunction with the visual stream. For instance, when provided with a video of a grassland featuring a girl and the dialogue 'Finally, I arrive at grandmother's ranch.' the model should generate an informative narration like 'A girl arrives at her grandmother's ranch.' Therefore, this necessitates the model to have the ability to concurrently decode dialogue and establish visual associations.

(ii) Event Dependency. Recognizing event dependencies is crucial for understanding video storylines, as these often contain fine-grained events with causal links between earlier and later events. For instance, in a sequence of movie clips [24], if an old man retrieves a flashlight from a car, it indicates his intention to illuminate something later. This type of dependency helps viewers construct a narrative bridge that connects past, present, and future elements in the video. Therefore, it is essential for the video model to manage not only individual events

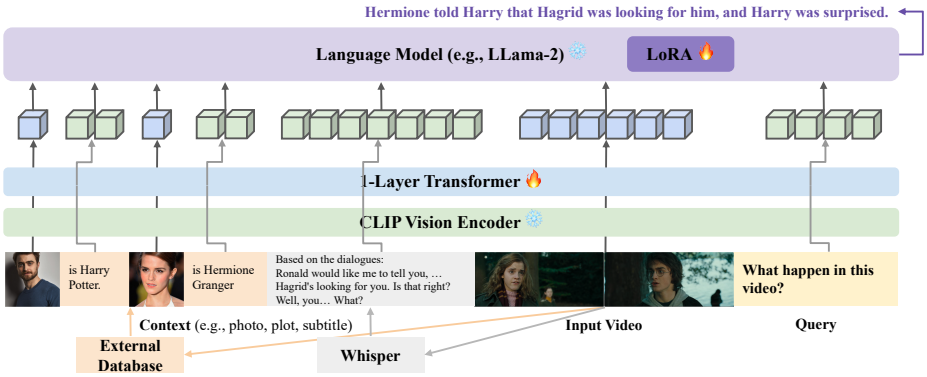


Fig. 3: Illustration of the pipeline of `MovieSeq`. Firstly, we embed the input video as an interleaved multimodal sequence (such as images, plots, videos, or subtitles), either by linking to an external database or leveraging annotations from offline models [6,63]. Then, we create an interleaved instruction (can be a combination of the above context) and feed it into the language model. The language model is trained to associate them and generate a comprehensive response.

but also to link multiple separate events, implying the need to jointly accept one or more video clip inputs and establish associations.

(iii) External Knowledge. Leveraging external information is often beneficial and necessary to improve the comprehension of videos, e.g., before watching a movie, viewers often find it helpful to access concise metadata like the movie's title, genre, plot summary, or glimpses of characters through trailers. It requires the assistant to flexibly integrate various external knowledge, e.g., characters' photos (visual) or movie plot documents (textual).

The above challenges are not exclusive to movies; they are widely applicable to other videos e.g., those social media videos usually involve additional contextual elements such as video titles, descriptions, user avatars, and ASR transcripts.

4 Methodology

We first illustrate `MovieSeq` pipeline and architecture, then provide more details of interleaved multimodal instruction construction.

4.1 Overview

Equip Video with Contexts. As shown in Fig.3(bottom), we first transform input videos as an interleaved multimodal sequence with additional context. For external knowledge, this includes character photos sourced from public datasets [23, 28] or textual movie plots from IMDb. For dialogues, we obtain speech transcriptions from offline models such as whisper [6,63], ensuring temporal alignment with the video.

Architecture. We illustrate our architecture in Fig. 3, building upon the common framework outlined in [49]. We utilize the pre-trained LLama-2 [75] as our language model, integrating LoRA [27] into all linear layers to achieve efficient

and effective tuning. We freeze all parameters except those in the input embedding layers, resulting in a mere 0.12% increase in trainable parameters by LoRA. The visual encoder is CLIP [62] ViT-B/16. To handle multiple images and videos of varying lengths, we introduce two modifications: (i) For visual inputs, we extract frame-wise `[CLS]` token instead of patch tokens. (ii) We project the visual embeddings through a single-layer transformer with temporal position encoding to capture temporal relationships. We recognize a performance constraint by not adopting patch tokens as in previous works [39,52]. While our focus is on exploring benefits in additional contexts. This enables efficient handling of multiple videos of varying lengths and image inputs, by treating images as a single frame, thereby enabling the flexibility of handling free-form interleaved inputs.

4.2 Interleaved Multi-Modal Instructions

Our primary focus is on leveraging available contexts to create an interleaved multi-modal instruction for narrative videos. To this end, we first introduce the generic instruction format, as detailed in Tab. 1, which comprises three terms: (i) context $\mathbf{X}_{\texttt{ctx}}$, which facilitates multi-modal interleaved context, encompassing elements like characters photos, dialogues, storylines, or video history; (ii) question $\mathbf{X}_{\texttt{q}}$, represents the user query, e.g., 'Can you help me describe this video?' or as a formulation of the downstream task, e.g., 'What is the relationship in this video?' (iii) answer $\mathbf{X}_{\texttt{a}}$ denotes the model's responses. This design allows the same question $\mathbf{X}_{\texttt{q}}$ to yield varied responses $\mathbf{X}_{\texttt{a}}$ under different contexts $\mathbf{X}_{\texttt{ctx}}$.

`USER :` $\mathbf{X}_{\texttt{ctx}}$ $\mathbf{X}_{\texttt{q}}$`\n MovieSeq:` $\mathbf{X}_{\texttt{a}}$ `<EOS>`

Table 1: Our interleaved multimodal instruction comprises the context $\mathbf{X}_{\texttt{ctx}}$, the question $\mathbf{X}_{\texttt{q}}$, and the answer $\mathbf{X}_{\texttt{a}}$. The context term provides necessary multimodal context (e.g., an arrangement of image, text, video, subtitles), and the model is trained to predict the answer $\mathbf{X}_{\texttt{a}}$ followed by `<EOS>` to indicate termination.

Video with Images. Considering the frequent needs for character identification [24,35,66] in movies, we exemplify our video with image instruction through this task. Practically, we identified two demands for character identification. Firstly, we need to determine which characters, whether one or several, appear in the video; Secondly, we want to interact with the video, supported by the identified character photos. Based on these requirements, we have devised the following two contexts and questions[1]:

(i_a) $\mathbf{X}_{\texttt{ctx}}$: There are several character photos: $\texttt{img}_1$ is $\texttt{name}_1$, $\texttt{img}_2$ is $\texttt{name}_2$, ..., $\texttt{img}_n$ is $\texttt{name}_n$ and a video `vid` .
$\mathbf{X}_{\texttt{q}}$: Who can be found in this video? If not, output None.

(i_b) $\mathbf{X}_{\texttt{ctx}}$: There have character photos: $\texttt{img}_1$ is $\texttt{name}_1$ (can have more) and a video `vid` .
$\mathbf{X}_{\texttt{q}}$: Please briefly describe this video (or other free-form query).

[1] We use grey color to represent the context e.g., `img` : image tokens; `vid` : video tokens; `sub` : subtitles; `plot` : selected movie plot.

With these two designs, we can identify who is in the video as well as generate a response with character identification.

Video with Plots. Movies come with a wealth of textual metadata, such as plots [72] or synopsis [28], which are valuable references for high-level comprehension. We chose MAD dataset [24, 67], and gathered movie plot synopsis from imdb website [2]. to fulfill our needs. However, the plot resembling a document tends to be quite long, and may surpass the context length of LLMs, accurately identifying relevant sections is crucial. We provide two solutions for plot sampling: **(a)** For instructions with a specific query such as a detailed question $\mathbf{X}_{\texttt{q}}$ in MovieQA [72]. We adopt retrieval-augmented approaches [22, 36]. We obtain sentence embeddings[3], perform sentence-paragraph retrieval using cosine similarities, and subsequently return the top-1 relevant paragraph as the context $\mathbf{X}_{\texttt{ctx}}$.

(b) For general queries such as clip captioning, we process each video clip centered at timestamp t in a movie with duration T by computing its ratio $r = \lfloor \frac{t}{T} \rfloor$. This ratio r is then utilized to pinpoint the r-th sentence in the plot and extract a paragraph using a sentence number as the window size w.

With the sampling plots, we define the instructions as:

(ii) $\mathbf{X}_{\texttt{ctx}}$: Based on the plot `plot` and the video `vid`.

where the $\mathbf{X}_{\texttt{q}}$: can be any free-form questions.

Video with Subtitles. Subtitles serve as a unique signal, offering extensive situation information, including human emotions, intentions, and specific terminology, and etc. We chose CMD [7] and TVC [35] as data sources, as each video clip in their datasets is paired with temporal aligned subtitles and captioning. This captioning, such as 'Beckett talks to Montgomery about her date and looks over at her coworkers', effectively links visuals with subtitles, making them suitable for supervising instruction-tuning. We design such a context:

(iii) $\mathbf{X}_{\texttt{ctx}}$: Based on the `sub` and the video `vid`.

Video with History. In long narrative videos, capturing temporal dependencies i.e., associations among multiple events is crucial. Leveraging historical videos as context, can enhance the understanding of the current or aid in predicting future events. For example, if a man is seen purchasing a train ticket, this suggests he may later appear at a station, thereby adding coherence to the narrative. MAD [67] is known for its long continuous narrations, which we employ to formulate the following instructions:

(iv) $\mathbf{X}_{\texttt{ctx}}$: There are $n+1$ video clips, ordered from the past to present: $\texttt{vid}_{\texttt{t-n}}$ $\texttt{cap}_{\texttt{t-n}}$ $\cdots$ $\texttt{vid}_{\texttt{t-2}}$ $\texttt{cap}_{\texttt{t-2}}$ $\texttt{vid}_{\texttt{t-1}}$ $\texttt{cap}_{\texttt{t-1}}$ $\texttt{vid}_{\texttt{t}}$.

In the above instruction, we include both the history video and their narrations $\texttt{vid}_{\texttt{t-i}}$ $\texttt{cap}_{\texttt{t-i}}$ as context together, which we find beneficial. The past narration can be derived from either annotations or the model's previous predictions in a recurrent inference setting.

[2] https://www.imdb.com/

[3] We use `all-mpnet-base-v2` from SentenceTransformers.

Datasets	#Sample	Video Len.	Task	Metric	Used Contexts
LVU [84]	1.5K	154.0s	Video Classification	Acc.	vid, sub
Movienet [28]	77K	key frames	Multi-label Cls.	F1-score	vid, img
MADv2 [24]	300K	4.1s-1.8h*	Audio Description	R-L, C, etc	vid, img, plot, hist.
CMD [7]	24K	132.0s	Video Retrieval	Recall@N	vid, sub
TVC [35]	174K	9.1s	Video Captioning	B4, R-L, etc.	vid, sub, img
MovieQA [72]	9.8K	202.7s	VideoQA	Acc.	vid, sub, plot

Table 2: Dataset statistics. Datasets are used for creating instruction and evaluation. These datasets vary in duration (ranging from keyframes to several seconds to minutes) and encompass diverse tasks with diverse contexts. The notation * indicates that each movie (avg 1.8h) comprises multiple short clips (avg. 4.1s), so that we can leverage history context (hist).

Notably, the above four instruction templates are designed for videos with one context. For videos with multiple contexts, such as MAD [24,67] and TVC [35], we expand the instructions by concatenating these contexts. Additionally, we utilize ChatGPT-3.5 to rephrase templates, aiming to achieve diversity and improve robustness. For example, the query: 'who can be found in this video?' with answer: `name` is transformed into a boolean question: 'Does `name` appear in this video?' with answer: 'Yes'.

4.3 Training Objectives

Despite the interleaved multimodal instruction designs, the training loss is computed based on $\mathbf{X}_{\mathtt{a}}$. We employ the common language modeling training objective, i.e., auto-regressive, as follows:

$$\max \sum_{i=1}^{L} \log p(\mathbf{X}_{\mathtt{a}} | \mathbf{X}_{\mathtt{ctx}}, \mathbf{X}_{\mathtt{q}}) = \prod_{i=1}^{L} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i | \mathbf{X}_{\mathtt{ctx}}, \mathbf{X}_{\mathtt{q}}, \mathbf{X}_{\mathtt{a},<i}), \tag{1}$$

where L the length of model response sequence i.e., $|\mathbf{X}_{\mathtt{a}}|$.

5 Experiments

In this section, we structure our experiments to investigate the following questions:
Q1. **Why LLM?** Without additional data, can `MovieSeq` exhibits flexibility and effectiveness compared with baselins across various setups?
Q2. **Why Context?** What impact does each video context have?
Q3. **Why Instruction-tuning?** Can we get a general model by multi-task multi-context instruction-tuning?
Q4. **Is Data Construction Trivial?** We perform ablation studies to examine setups in instruction construction, such as template design, sampling strategy.

5.1 Datasets and Settings

We assess `MovieSeq` on six datasets spanning various tasks, as summarized in Tab. 2. For each setting, we briefly introduce the dataset and evaluation metrics. For non-generative tasks e.g., video classification (i,ii) and video-text retrieval (iv),

we express the adaptation of our generative language model to these predictions, with further details in the Supp.

(i) LVU [84] comprises 30K videos from 3K movies, focusing on content understanding (relationship, speaking style, scene), metadata prediction (e.g., director, writer, year), and user engagement (e.g., Youtube like ratio). As `MovieSeq` is a generative model, we concentrate on content understanding, which has a clear, inferable interpretation. We employ Whisper [63] to extract subtitles for LVU videos. The metric employed is the top-1 accuracy. Our `MovieSeq` addresses this classification task by directly generating the category names.

(ii) Movienet [28] is a movie dataset that offers keyframes and character mappings for each frame. We adhere to the settings described in [23] to examine the character identification, a multi-label classification task. Our `MovieSeq` addresses this classification task by directly generating character names without the need for score thresholds. Thus, we opt for Recall, Precision, and F1-score as our evaluation metrics.

(iii) MADv2 [24] comprises a collection of 264K movie audio descriptions. Its goal is to automatically generating narrations that align with individual video elements, incorporate character names, and ensuring temporal coherence. Following the official settings [23,24], we evaluate predictions against groundtruth using ROUGE-L [43] (R-L), CIDEr [76] (C), SPICE [3] (S), and Recall@k within N neighbours (R@k/N) [23] based on BertScores [93]. Notably, the above AD metric focuses on single captions and overlooks correlations among multiple sentences. As suggested by [34], we employ Repetition@4 (Rep@4) to evaluate sentence repetition in our ablation studies.

(iv) CMD [7], a long-range video-text retrieval benchmark, includes key scenes from over 3K movies, each paired by a high-level description, such as the movie's storyline. We align with the recent baseline [70], and adopt the Geometric Mean, R@1, R@5, and R@10 as metrics. Although video-text retrieval is a contrastive task, we transform it by generating captions for each video. Then, we conduct sentence retrieval using these generated captions against ground-truth queries.

(v) TVC [35] is a television captioning dataset, consisting of 262K descriptions matched with 108K video segments. TVC captions not only describe visual elements of the content but also incorporate dialogue elements from the subtitles. We follow [45] and adopt BLEU4 [60] (B4), METEOR [10] (M), ROUGE-L [43] (R-L), and CIDER [76] (C) as mertics.

(vi) MovieQA [72], a video question-answering dataset, presents a challenge for the model to answer reasoned questions. Each question offers five multiple-choice options, with accuracy serving as the evaluation metric.

Implementation Details. For MAD, we sample 8 frames per clip. For Movienet, we sample each available frame. For CMD, TVC, and MovieQA, we sample 32 frames per video. For LVU, we sample 64 frames per video. We train our model for 20 epochs on each dataset, using a learning rate of 3e-5. For LoRA [27] settings, we use a rank of 16 with an alpha of 16. For subtitles or plots, the maximum length is set to 512. For model responses, the default maximum length is set to 64.

Methods	Style	Relation	Speak	Scene	**Avg.**
VideoBERT [84]	Discr.	52.8	37.9	54.9	48.5
Obj. Tran. [84]	Discr.	53.1	39.4	56.9	49.8
VIS4mer [29]	Discr.	57.1	40.8	67.4	56.9
LF-VILA [84]	Discr.	61.5	41.3	68.0	60.9
S5 [81]	Discr.	67.1	42.1	73.5	67.0
MA-LMM [26]	Gen.	58.2	44.8	80.3	61.1
`MovieSeq` (vid)	Gen.	61.0	37.6	**76.8**	58.5
`MovieSeq` (sub)	Gen.	61.0	61.0	62.4	61.5
`MovieSeq` (vid, sub)	Gen.	**75.6**$_{\uparrow 14.6}$	**63.0**$_{\uparrow 25.4}$	65.9$_{\downarrow 10.9}$	**68.1**$_{\uparrow 9.5}$

Table 3: Video Classification on LVU [84] content understanding. Discr. means discriminative model e.g., classifier. Gen. means generative language model. The color green and red denote the gain or drop by the introduction of subtitles compared with the variant (vid).

5.2 Main Results (Q1&Q2).

In this section, we compare `MovieSeq` on each benchmark individually, showcasing its flexibility and effectiveness relative to baseline methods. *Notably, for a fair comparison, we report results using individual training set.* We will study the effect of multi-task instruction-training in ablation section.

LVU. In Tab. 3, we initially evaluate `MovieSeq`, on the LVU content understanding test set. This is primarily a video classification task and no baseline has attempted to introduce subtitles. We observe that the dialogue significantly aids in assessing relations (+14.6%) and speech (+25.4%), while it diminishes the scene understanding (-10.9%). This makes sense as dialogue usually reflects the situational content but does not directly benefit the purely visual aspects.

Methods	Style	Precision	Recall	F1-score
CLIP cos. [62]	Discr.	39.6	71.8	51.0
TFM Decoder [23]	Discr.	75.9	**82.7**	79.1
`MovieSeq`	Gen.	**88.5**	75.5	**81.4**

Table 4: Actor Identification on Movienet [28].

Movienet. In Tab. 4, we evaluate character identification on Movienet, a multi-label classification task. The baseline results are sourced from [23], where we select their score threshold by yielding the highest F1-score. It's worth noting that this task is a multi-label classification task. Our method surpasses the baseline TDM Decoder [23], a module specifically tailored for this task, by 2.3 in F1-score. With Tab. 3 and Tab. 4, we demonstrate that the generative language model can perform well as a discriminative model.

Methods	Pretraining Data	R-L	C	S	R@5/16
ClipCap [56]	CC3M	8.5	4.4	1.1	36.5
CapDec* [57]	AV-AD	8.2	6.7	1.4	-
AutoAD-I [24]	-	9.3	6.7	2.4	-
AutoAD-I [24]	AV-AD & WebVid	11.9	14.3	4.4	42.1
AutoAD-II [23]	-	12.7	18.3	-	45.6
AutoAD-II [23]	AV-AD & WebVid	13.4	19.5	-	50.8
`MovieSeq` (vid,img,plot,hist.)		**15.5**	**24.4**	**7.0**	**51.6**

Table 5: Audio Description generation task on MAD-v2. The baseline results are from AutoAD-I&II [23, 24]. Please see Tab.9 for clear ablation of each context.

Methods	#PT Data	#Frame	Geo. Mean	R@1	R@5	R@10
MoEE [55]	-	-	5.9	1.9	7.8	13.4
TeachText [16]	-	-	23.2	12.1	27.4	37.5
Frozen [8]	5M	32	-	12.6	28.4	36.3
LF-VILA [70]	8M	32	26.4	13.6	32.5	41.8
VINDLU [70]	25M	32	-	18.4	36.4	44.3
TESTA [70]	5M	32	-	21.5	42.4	**50.7**
`MovieSeq` (vid)	-	32	9.2	4.3	11.8	14.9
`MovieSeq` (sub)	-	32	33.1	21.2	38.0	44.8
`MovieSeq` (vid, sub)	-	32	**38.9**$_{\uparrow 29.7}$	**25.8**$_{\uparrow 21.5}$	**45.3**$_{\uparrow 33.5}$	50.3$_{\uparrow 35.4}$

Table 6: Video-Text Retrieval on CMD [7] test set. The color green denotes the gain by the subtitles context compared with the variant (vid).

MAD. In Tab. 5, we display the results on the MADv2 audio description benchmark. Notably, this task requires identifying character names for each clip. Thus, we reuse the model trained on Tab. 4 to predict character names. Our model, `MovieSeq`, without additional pretraining, achieves competitive results over the best baseline, i.e., a 4.9% higher CIDEr.

CMD. We carry out experiments on the CMD video-text retrieval task in Tab. 6. In this task, we unexpectedly find that `MovieSeq`(vid) variant performs significantly poorly. Upon analyzing their predictions, we find that: CMD text queries such as 'Max flies the drone home before his dad notices it's missing.' focus on the key scene storyline, including names, relationships, and different focal points. The video-only generative tuning fails to find relevant cues to derive reasonable captions, resulting in a lot of hallucinations. However, with the aid of dialogues, we significantly boost the performance, achieving a 29.74 Geo. Mean. improvement. This demonstrates that for a generative model, the quality of the tuning data is important (i.e., the answer should be derivable from the context).

TVC. In Tab. 7, we validate `MovieSeq` on the video captioning task, TVC. This benchmark originally provides subtitles as inputs. Without pretraining, `MovieSeq` outperforms several baselines, including those utilizing subtitles such as HERO [40] and VALUE [41]. This improvement demonstrates that language models with reasoning provide better visual-dialogue association. Additionally, we found that the inclusion of character images does not significantly enhance performance. This could be because, in TVC, the subtitles already provide the names (e.g., 'Chandler: Yeah, well...'), and TVC only contains six TV series, thus having a limited number of characters, most are seen during training.

Methods	Context	PT Data	Frame	B4	R-L	M	C
MMT [35]	vid, sub	-	3 FPS	10.8	32.8	16.9	45.3
HERO [40]	vid, sub	7.6M	2/3 FPS	12.3	34.1	17.6	49.9
VALUE [41]	vid, sub	-	2/3 FPS	11.6	33.9	17.6	50.5
SWINBERT [45]	vid	-	64	14.5	36.1	18.5	55.4
GIT$_B$ [80]	vid	14M	6	13.0	33.2	16.6	47.3
All-in-one [80]	vid	105M	3	12.5	-	**20.4**	56.3
`MovieSeq`	vid	-	32	13.7	34.0	17.4	50.9
`MovieSeq`	sub	-	32	11.3	31.0	14.8	42.7
`MovieSeq`	vid, sub	-	32	17.5	37.5	19.4	63.5
`MovieSeq`	vid, sub, img†	-	32	**17.9**$_{\uparrow 4.2}$	**38.1**$_{\uparrow 4.1}$	19.9$_{\uparrow 2.5}$	**64.8**$_{\uparrow 13.9}$

Table 7: Video Captioning on TVC [35] val set. Img† means the character's photos we collected. The color green denotes the gain by the subtitles and character photos compared with the variant (vid).

MovieQA. In Tab. 8, we validate `MovieSeq` on the video question-answering task, MovieQA. This is a challenging setting, as the questions often contain types like 'How' and 'Why', which require reasoning beyond video. Thus, we observe that visual-only baselines yield weak performance (only 27.32 Acc.). We notice the gains of plots as they offer distilled information. But our goal is not to exceed baseline scores via plot usage, but to offer new way for understanding video context within existing LLM methods, and show how to identify valuable parts from document-level plots. This will be studied in ablation Fig.4(b).

Methods	Acc.
SSCB [72] (vid, subs)	34.20
PAMN [32] (vid, subs)	43.34
HMMN [79] (vid, subs)	46.28
SSCB [72] (plot)	56.70
`MovieSeq` (vid)	27.32
`MovieSeq` (subs)	27.10
`MovieSeq` (plot)	63.54
`MovieSeq` (vid, subs)	48.32$_{\uparrow 21.9}$
`MovieSeq` (vid, plot)	**66.41**$_{\uparrow 39.1}$

Table 8: Video QA on MovieQA [72] val set.

5.3 Ablation Studies

Effect of Multimodal Contexts (Q2). In most previous tables, we have presented variants to ablate the impact of different modality contexts. In Tab. 9, we study the effect of different contexts (images, plot, history) for the audio description task. There are several findings: **(i)** From row 2, the external plot as a reference greatly improves the performance (+4.7 CIDEr). **(ii)** From rows 3-4, with the past videos (with their narrations) as context, the model can effectively avoid repetition. If the past context from prediction (recurrent), it does not influence the original AD metrics. However, if the past context is sourced from the ground-truth (oracle), it can greatly boost them. This suggests that a reliable past context is more helpful. **(iii)** Rows 5 and 6 in the table showcase distinct approaches: row 5 employs CLIP vision scores for character retrieval, whereas row 6 is the model trained with MovieNet [28]. This results indicate that character identification from additional instructions is necessary. Lastly, with the above designs combined, we can jointly boost the AD metrics as well as avoid repetition.

Rows	Settings	Context	AD metrics				Multi-sentences
			R-L	C	S	R@5/16	Rep@4 (↓)
1	Video	vid	10.6	11.4	2.5	48.6	1.33
2	w/ Plot	vid, plot	12.9	16.9$_{\uparrow 5.5}$	4.4	48.8	0.48$_{\downarrow 0.85}$
3	w/ Past	vid, hist.	10.4	11.8$_{\uparrow 1.5}$	2.6	50.1	0.11$_{\downarrow 10.4}$
5	w/ Characters (clip)	vid, img	13.1	19.7	5.9	47.3	1.30
6	w/ Characters (movienet)	vid, img	15.2	23.6$_{\uparrow 12.2}$	**7.2**	50.3	0.97$_{\downarrow 0.36}$
7	`MovieSeq`	vid, img, plot, hist.	**15.5**	**24.4**$_{\uparrow 13.0}$	7.0	**51.6**	0.12$_{\downarrow 12.1}$

Table 9: Ablative Studies of different context on MAD. The color green denotes the gain by the subtitles and character photos compared with the variant (vid). For brevity, we only highlight the (C)IDER and Rep@4.

Effect of Multi-Task Co-Training (Q3). In Tab. 10, we investigate the effect of multi-task co-training for a more robust model. Due to notable data imbalances across the datasets such as MAD being 200x larger than LVU. We constructed a development set by selecting 1K samples from each dataset, including MAD (audio desc.), MovieNet (vid. cls.), and TVC (vid. cap.), to ensure comparable sizes with different tasks.

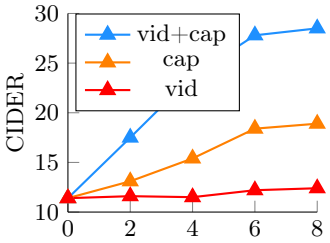
Multi-task co-training results in an average 1.5 improvement across three tasks, demonstrating that multi-task learning enhances individual capabilities. This highlights the language model's ability to acquire commonsense across diverse objectives and contexts.

Training strategy	Multi -task?	MAD (CIDEr)	MovieNet (F1-score)	TVC (CIDEr)	**Avg**
individual	✗	22.5	78.6	60.2	53.8
co-training	✓	23.8	79.2	62.7	55.2$_{\uparrow 1.5}$

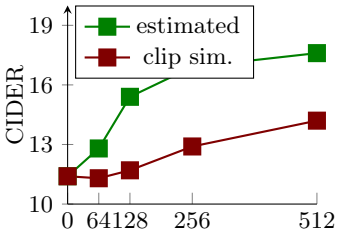
Table 10: Effect of multi-task co-training. These three datasets are similar in scale but encompass different tasks.

Strategy of Data Construction (Q4). In 'Video with History' setting, the template design and the history clip number play an important factor. Ablation studies conducted on the MAD benchmarks with oracle history narrations, as illustrated in Fig. 4(a), revealed: **(i)** With the same number of clips, the ▲vid+cap variant surpasses ▲cap alone, whereas ▲vid exhibits negligible change. This emphasizes the significance of narrations in grasping event relationships and the challenge of reasoning in vision solely. **(ii)** Increasing the number of input clips consistently leads to performance gains in the 'vid+cap' and 'cap' variants.

In 'Video with Plots' setting, our temporal ■estiamted strategy is contrasted with a variant that utilizes ■ clip vision-text similarities for filtering, as shown in Fig. 4(b). Our approach results highlighting the necessary of a reliable plot sampling method. Additionally, we observe a consistent trend of improved performance with increased context length, emphasizing the necessity of developing language models capable of processing longer contexts.



(a) Diff. # of history clip under diff. instruction template (iv).

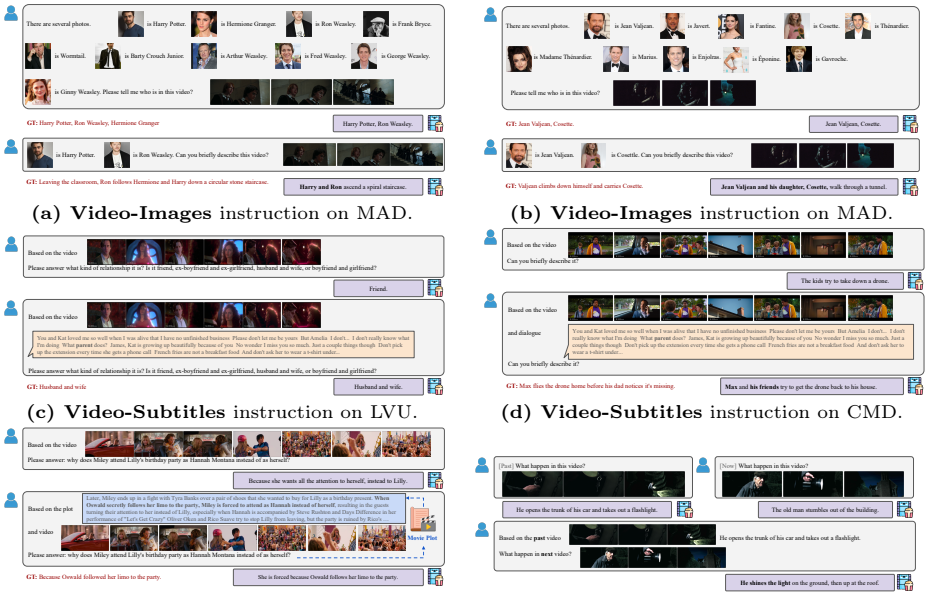


(b) Diff. input plot length (# of tokens) under diff. strategy.

Fig. 4: Effect of Data Construction on MAD dataset.

5.4 Visualization.

In Fig. 5, we provide visualization of `MovieSeq` under different kinds of interleaved multimodal instruction. **(a-b) Video with Images.** First, we provide the top-10 character photos and let the model decide who is present. Next, using these selected characters as a context, we can generate an accurate narration with character names. `MovieSeq` can handle cases with multiple characters. **(c-d) Video with Subtitles.** We use subtitles to enhance the situational understanding. As seen in (c), it remains challenging to determine whether the relationship is that of friends or a couple from visual. Notably, as shown in (d), with the dialogue, the language model can associate the name (i.e., Max) with relationships (i.e., friends) and produce a meaningful caption. **(e) Video with Plots:** We sample



(a) **Video-Images** instruction on MAD.

(b) **Video-Images** instruction on MAD.

(c) **Video-Subtitles** instruction on LVU.

(d) **Video-Subtitles** instruction on CMD.

(e) **Video-Plots** instruction on MovieQA.

(f) **Video-History** instruction on MAD.

Fig. 5: Visualization of `MovieSeq` by providing different kinds of interleaved multimodal prompts for different applications.

an example from MovieQA. As shown, questions such as 'how' are primarily challenging to derive from the visual stream. Therefore, by using a retrieval-augmented plot as a reference (e.g., the background of why Miley attends the birthday party as Hannah Montana), the model can find the cue and derive the correct reason. **(f) Video with History:** In the multi-video settings, we use it to enhance the association of different events. As shown in (f), with the past prediction as context (e.g., "He opens the trunk of his car and takes out a flashlight"), the model can derive the next narration more naturally (e.g., "shines the light since he takes out a flashlight in the previous scene").

6 Conclusion and Limitations

In this paper, we introduce `MovieSeq`, a video-language model designed for video context understanding. By modeling narrative videos as an interleaved multimodal sequence with external knowledge and additional modalities, `MovieSeq` can support flexible interleaved multimodal instruction, and resolve several challenges including dialogues understanding, character identification, and event dependency. Furthermore, we illustrate how to collect the corresponding tuning data for each type of interleaved prompt. We demonstrate the effectiveness and flexibility of our `MovieSeq` model across six diverse datasets.

Despite its strengths, our model still has *limitations*. We have not yet incorporated novel architectural designs, instead choosing to restrict visual branches (e.g., using CLS tokens instead of patch tokens) to save token length. Exploring ways to effectively model long context without affecting visual input remains a topic worthy of discussion. We leave them in our future work.

Acknowledgements. This project is supported by the DSO National Laboratories. We extend our gratitude to Mattia Soldan for his assistance with MAD visualization examples, to Tengda Han's help with providing character identification modules and data, and to Makarand Tapaswi for his help on the MovieQA datasets.

References

1. Adel Ahmadyan, Liangkai Zhang, A.A.J.W.M.G.: Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14. pp. 382–398. Springer (2016)
4. Argaw, D.M., Lee, J.Y., Woodson, M., Kweon, I.S., Heilbron, F.C.: Long-range multimodal pretraining for movie understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13392–13403 (2023)
5. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., et al.: Openflamingo: An open-source framework for training large autoregressive vision-language models. arXiv preprint arXiv:2308.01390 (2023)
6. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio (2023)
7. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: Proceedings of the Asian Conference on Computer Vision (2020)
8. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1728–1738 (2021)
9. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: A clip-hitchhiker's guide to long video retrieval. arXiv preprint arXiv:2205.08508 (2022)
10. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005)
11. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: ICML. vol. 2, p. 4 (2021)
12. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. NeurIPS pp. 1877–1901 (2020)
13. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)

14. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: Videollm-online: Online video large language model for streaming video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18407–18418 (2024)
15. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
16. Croitoru, I., Bogolin, S.V., Leordeanu, M., Jin, H., Zisserman, A., Albanie, S., Liu, Y.: Teachtext: Crossmodal generalized distillation for text-video retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11583–11593 (2021)
17. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2305.06500 (2023)
18. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
19. Gao, D., Ji, L., Zhou, L., Lin, K.Q., Chen, J., Fan, Z., Shou, M.Z.: Assistgpt: A general multi-modal assistant that can plan, execute, inspect, and learn. arXiv:2306.08640 (2023)
20. Gao, D., Zhou, L., Ji, L., Zhu, L., Yang, Y., Shou, M.Z.: Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14773–14783 (2023)
21. Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: Proceedings of the IEEE international conference on computer vision. pp. 5842–5850 (2017)
22. Guu, K., Lee, K., Tung, Z., Pasupat, P., Chang, M.: Retrieval augmented language model pre-training. In: International conference on machine learning. pp. 3929–3938. PMLR (2020)
23. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad ii: The sequel-who, when, and what in movie audio description. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13645–13655 (2023)
24. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: AutoAD: Movie description in context. In: CVPR (2023)
25. Han, T., Xie, W., Zisserman, A.: Temporal alignment networks for long-term video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2906–2916 (2022)
26. He, B., Li, H., Jang, Y.K., Jia, M., Cao, X., Shah, A., Shrivastava, A., Lim, S.N.: Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13504–13514 (2024)
27. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
28. Huang, Q., Xiong, Y., Rao, A., Wang, J., Lin, D.: 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. pp. 709–727. Springer (2020)

29. Islam, M.M., Bertasius, G.: Long movie clip classification with state-space video models. In: European Conference on Computer Vision. pp. 87–104. Springer (2022)
30. Islam, M.M., Hasan, M., Athrey, K.S., Braskich, T., Bertasius, G.: Efficient movie scene detection using state-space transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18749–18758 (2023)
31. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
32. Kim, J., Ma, M., Kim, K., Kim, S., Yoo, C.D.: Progressive attention memory network for movie story question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8337–8346 (2019)
33. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: a large video database for human motion recognition. In: 2011 International conference on computer vision. pp. 2556–2563. IEEE (2011)
34. Lei, J., Wang, L., Shen, Y., Yu, D., Berg, T.L., Bansal, M.: Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. arXiv preprint arXiv:2005.05402 (2020)
35. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
36. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in Neural Information Processing Systems **33**, 9459–9474 (2020)
37. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
38. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
39. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
40. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200 (2020)
41. Li, L., Lei, J., Gan, Z., Yu, L., Chen, Y.C., Pillai, R., Cheng, Y., Zhou, L., Wang, X.E., Wang, W.Y., et al.: Value: A multi-task benchmark for video-and-language understanding evaluation. arXiv preprint arXiv:2106.04632 (2021)
42. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. arXiv preprint arXiv:2311.17043 (2023)
43. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
44. Lin, K., Ahmed, F., Li, L., Lin, C.C., Azarnasab, E., Yang, Z., Wang, J., Liang, L., Liu, Z., Lu, Y., et al.: Mm-vid: Advancing video understanding with gpt-4v (ision). arXiv preprint arXiv:2310.19773 (2023)
45. Lin, K., Li, L., Lin, C.C., Ahmed, F., Gan, Z., Liu, Z., Lu, Y., Wang, L.: Swinbert: End-to-end transformers with sparse attention for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17949–17958 (2022)
46. Lin, K.Q., Wang, J., Soldan, M., Wray, M., Yan, R., Xu, E.Z., Gao, D., Tu, R.C., Zhao, W., Kong, W., et al.: Egocentric video-language pretraining. Advances in Neural Information Processing Systems **35**, 7575–7586 (2022)

47. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2794–2804 (2023)
48. Lin, Y.B., Lei, J., Bansal, M., Bertasius, G.: Eclipse: Efficient long-range video retrieval using sight and sound. In: European Conference on Computer Vision. pp. 413–430. Springer (2022)
49. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
50. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3042–3051 (2022)
51. Lu, Y., Li, C., Liu, H., Yang, J., Gao, J., Shen, Y.: An empirical study of scaling instruct-tuned large multimodal models. arXiv:2309.09958 (2023)
52. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
53. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. arXiv preprint arXiv:2308.09126 (2023)
54. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9879–9889 (2020)
55. Miech, A., Laptev, I., Sivic, J.: Learning a text-video embedding from incomplete and heterogeneous data. arXiv preprint arXiv:1804.02516 (2018)
56. Mokady, R., Hertz, A., Bermano, A.H.: ClipCap: CLIP prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
57. Nukrai, D., Mokady, R., Globerson, A.: Text-only training for image captioning using noise-injected CLIP. arXiv preprint arXiv:2211.00575 (2022)
58. OpenAI: Gpt-4 technical report (2023)
59. OpenAI: Introducing chatgpt. https://openai.com/blog/chatgpt/ (2023)
60. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
61. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 724–732 (2016)
62. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) ICML. pp. 8748–8763 (2021)
63. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)
64. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog p. 9 (2019)

65. Real, E., Shlens, J., Mazzocchi, S., Pan, X., Vanhoucke, V.: Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5296–5305 (2017)
66. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015)
67. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5026–5035 (2022)
68. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., et al.: Moviechat: From dense token to sparse memory for long video understanding. arXiv preprint arXiv:2307.16449 (2023)
69. Srivastava, D., Singh, A.K., Tapaswi, M.: How you feelin'? learning emotions and mental states in movie scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2517–2528 (2023)
70. Sun, Y., Xue, H., Song, R., Liu, B., Yang, H., Fu, J.: Long-form video-language pre-training with multimodal temporal contrastive learning. Advances in neural information processing systems **35**, 38032–38045 (2022)
71. Surís, D., Menon, S., Vondrick, C.: Vipergpt: Visual inference via python execution for reasoning. arXiv:2303.08128 (2023)
72. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
73. Tong, Z., Song, Y., Wang, J., Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. Advances in neural information processing systems **35**, 10078–10093 (2022)
74. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
75. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton-Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288 (2023)
76. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
77. Vicol, P., Tapaswi, M., Castrejon, L., Fidler, S.: Moviegraphs: Towards understanding human-centric situations from videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8581–8590 (2018)
78. Wang, A.J., Li, L., Lin, K.Q., Wang, J., Lin, K., Yang, Z., Wang, L., Shou, M.Z.: Cosmo: Contrastive streamlined multimodal model with interleaved pre-training. arXiv preprint arXiv:2401.00849 (2024)

79. Wang, A., Luu, A.T., Foo, C.S., Zhu, H., Tay, Y., Chandrasekhar, V.: Holistic multi-modal memory network for movie question answering. IEEE Transactions on Image Processing **29**, 489–499 (2019)
80. Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K.Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., et al.: All in one: Exploring unified video-language pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6598–6608 (2023)
81. Wang, J., Zhu, W., Wang, P., Yu, X., Liu, L., Omar, M., Hamid, R.: Selective structured state-spaces for long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6387–6397 (2023)
82. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., et al.: Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191 (2022)
83. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 284–293 (2019)
84. Wu, C.Y., Krahenbuhl, P.: Towards long-form video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1884–1894 (2021)
85. Wu, C.Y., Li, Y., Mangalam, K., Fan, H., Xiong, B., Malik, J., Feichtenhofer, C.: Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13587–13597 (2022)
86. Xie, J., Feng, J., Tian, Z., Lin, K.Q., Huang, Y., Xia, X., Gong, N., Zuo, X., Yang, J., Zheng, Y., et al.: Learning long-form video prior via generative pre-training. arXiv preprint arXiv:2404.15909 (2024)
87. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Zero-shot video question answering via frozen bidirectional language models. Advances in Neural Information Processing Systems **35**, 124–141 (2022)
88. Yang, A., Nagrani, A., Laptev, I., Sivic, J., Schmid, C.: Vidchapters-7m: Video chapters at scale. arXiv preprint arXiv:2309.13952 (2023)
89. Yang, Z., Li, L., Wang, J., Lin, K., Azarnasab, E., Ahmed, F., Liu, Z., Liu, C., Zeng, M., Wang, L.: Mm-react: Prompting chatgpt for multimodal reasoning and action. arXiv:2303.11381 (2023)
90. Zellers, R., Lu, J., Lu, X., Yu, Y., Zhao, Y., Salehi, M., Kusupati, A., Hessel, J., Farhadi, A., Choi, Y.: Merlot reserve: Neural script knowledge through vision and language and sound. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16375–16387 (2022)
91. Zhang, C., Lin, K., Yang, Z., Wang, J., Li, L., Lin, C.C., Liu, Z., Wang, L.: Mm-narrator: Narrating long-form videos with multimodal in-context learning. arXiv preprint arXiv:2311.17435 (2023)
92. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv:2307.03601 (2023)
93. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675 (2019)
94. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)