Multilingual Synopses of Movie Narratives: A Dataset for Vision-Language Story Understanding

Yidan Sun[1], Jianfei Yu[2], and Boyang Li[1]

[1]College of Computing and Data Science, Nanyang Technological University
[2]School of Computer Science and Engineering, Nanjing University of Science and Technology
SUNY0053@e.ntu.edu.sg, jfyu@njust.edu.cn, boyang.li@ntu.edu.sg

Abstract

Story video-text alignment, a core task in computational story understanding, aims to align video clips with corresponding sentences in their descriptions. However, progress on the task has been held back by the scarcity of manually annotated video-text correspondence and the heavy concentration on English narrations of Hollywood movies. To address these issues, in this paper, we construct a large-scale multilingual video story dataset named Multilingual Synopses of Movie Narratives (M-SYMON), containing 13,166 movie summary videos from 7 languages, as well as manual annotation of fine-grained video-text correspondences for 101.5 hours of video. Training on the human annotated data from M-SYMON outperforms the SOTA methods by 15.7 and 16.2 percentage points on Clip Accuracy and Sentence IoU scores, respectively, demonstrating the effectiveness of the annotations. As benchmarks for future research, we create 6 baseline approaches with different multilingual training strategies, compare their performance in both intra-lingual and cross-lingual setups, exemplifying the challenges of multilingual video-text alignment. The dataset is released at: https://github.com/insundaycathy/M-SyMoN

1 Introduction

Computational story understanding aims to empower AI systems with the ability to delve into the intricacies of diverse stories, unlocking their deep semantics such as character motivations and intentions (Emelin et al., 2021; Rashkin et al., 2018), event structures (Chambers and Jurafsky, 2008; Du et al., 2021), and social relationships among story characters (Elson et al., 2010; Chaturvedi et al., 2016; Kim and Klinger, 2019). In recent years, computation story understanding has garnered significant research interest (Dong et al., 2023; Sang et al., 2022; Andrus et al., 2022; Xu et al., 2022; Han et al., 2023) and many story understanding tasks (Wu and Krahenbuhl, 2021; Choi et al., 2021; Gu et al., 2023) have emerged.

Story video-text alignment is a fundamental task of computational story understanding, which aims to find the best correspondence between a sequence of video clips and a sequence of sentences (see an alignment example in Figure 1). Different from traditional video-text retrieval that relies on keyword or temporal cue matching (Wang et al., 2021b), story video-text alignment requires various story understanding abilities such as causal chain reasoning, mental state description, and long-range context understanding (Sun et al., 2022). Establishing such video-text correspondence will facilitate many applications such as text-to-video generation (Liu et al., 2019; Balaji et al., 2019), and visual story generation (Huang et al., 2019). Thus, the story video-text alignment task has recently attracted increasing attention (Dogan et al., 2018; Wang et al., 2021a; Sun et al., 2024).

However, a major obstacle of this task stems from the scarcity of annotated data, as it is costly and time-consuming to manually annotate sentence-level alignments between text and clip sequences. The story understanding datasets used in previous studies are limited for several reasons. First, although the Large Scale Movie Description Challenge (LSMDC) dataset (Rohrbach et al., 2017) manually aligns 158 hours of movies and audio descriptions, it only considers one-to-one matching between movie clips and audio descriptions and its audio descriptions are designed for visually impaired audiences, containing excessive details. Second, the YouTube Movie Summary (YMS) dataset (Dogan et al., 2018) is small and only contains 6.7 hours of video and texts, and thus is typically employed for video-text alignment evaluation. Lastly, Hollywood movies and English narrations are dominant in existing story understanding datasets (Huang et al., 2020; Soldan et al., 2022; Sun et al., 2022; Lu et al., 2023), neglecting the

Findings of the Association for Computational Linguistics: EACL 2024, pages 13488–13504

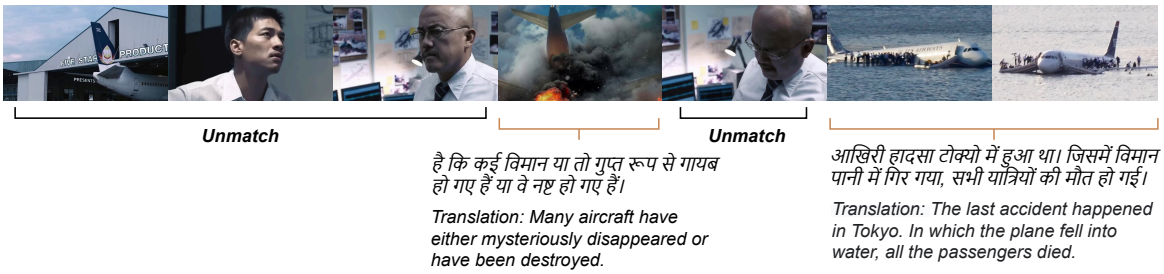

Figure 1: An example alignment between a video clip sequence and a sentence sequence from M-SYMON. One text chuck may correspond to several video clips, while some video clips may not match any textual description. The example is from https://youtu.be/n5v9hzSYxPQ, a summary of the movie *407 Dark Flight* (2012)

.

importance of language and cultural diversity.

To address the aforementioned limitations, we construct a multilingual video story dataset named Multilingual Synopses of Movie Narratives (M-SYMON). M-SYMON is sourced from movie recap videos on YouTube and contains 13,166 videos spanning 7 languages and totalling 2,136 hours. Furthermore, we manually annotated a portion of this dataset, providing exact video-text correspondence for 480 videos or 101.5 video hours. Compared to LSMDC and YMS, our annotated video-text alignment subset is large-scale, contains one-to-many matching and unmatched text or video clips, and covers 7 different languages.

We further investigate the multi-lingual characteristics of the dataset, and make the following observations. First, in the intra-lingual setup, compared to translating all languages to English for training and inference, additional language-specific finetuning on weakly supervised data brings an average improvement of 5.4 percentage points across 7 languages. Adding a small portion of the manual annotations further boosts performance. Second, in the cross-lingual setup, for source and target languages that are linguistically similar (e.g., Spanish to Portuguese or French), the transfer performance is generally good; for source and target languages that are different (e.g., Chinese to English or Hindi), the transfer performance is quite limited. Third, an out-of-domain evaluation on the YMS dataset shows that training on the weakly supervised data from M-SYMON outperforms the state of the art methods by 12.3 and 13.2 percentage points on Clip Accuracy and Sentence IoU scores. Moreover, adding manually annotated video-text alignment data further improves the performance by 2.4 and 3.0 percentage points, indicating the utility of our annotated alignment data.

The main contributions of this work are summarized as follows:

- We construct a large-scale multilingual video story understanding dataset named M-SYMON, containing 13,166 videos in 7 languages and totaling 2,136 hours.
- We manually annotate the fine-grained alignment between video clips and sentences of 480 videos for a total of 101.5 hours.[1]
- We create a number of multilingual video-text alignment methods to benchmark the M-SYMON dataset. Extensive results on both M-SYMON and YMS demonstrate the significance of our multilingual dataset and the utility of the human video-text alignment annotations.

2 Related Work

2.1 Story Video-Text Alignment

Story video-text alignment involves aligning sequences of video clips, typically from movies, with text captions (Cour et al., 2008; Tapaswi et al., 2015; Wang et al., 2021a; Dogan et al., 2018). The common approach is to first learn a video-text similarity metric and then calculate the alignment using dynamic programming (e.g., DTW) (Zhang et al., 2023b; Dvornik et al., 2021). Recent methods encode videos with 3D-convolution (Sun et al., 2024) or ViT (Li et al., 2023) and texts with Transformer models (Han et al., 2022; Li et al., 2023; Zhang et al., 2023a), then optimize alignment via soft-DTW (Zhang et al., 2023a; Han et al., 2022) or calculate the alignment from video-text similarity (Zhang et al., 2023b).

[1]The video URLs and video-text alignment annotations in M-SYMON will be made publicly available.

Due to the lack of sentence-level video-text alignment annotations, most models are trained with timestamp-based weak supervision. Although the LSMDC dataset (Rohrbach et al., 2017) contains a large amount of manual annotation, its overly fine-grained text data interferes with event understanding (Sun et al., 2022); it also lacks one-to-many matching and unmatched items. The YMS dataset (Dogan et al., 2018) contains only 6.7 hours of movie summaries with exact video-text alignment and cannot serve as a comprehensive test benchmark.

2.2 Movie Story Datasets

Movies are a popular source of video-text story content. LSMDC (Rohrbach et al., 2017) and Movie Audio Description (MAD) (Soldan et al., 2022) provide movie clips with audio descriptions for the visually impaired. Although the audio descriptions are accurate, they deviate significantly from realistic styles of story narration. The Condensed Movie Descriptions (CMD) dataset (Bain et al., 2020) offers 7 to 11 key clips per movie with one-sentence descriptions. The Pororo dataset (Kim et al., 2017) includes 20-minute cartoon episodes with in-show conversations and human-written descriptions. Although CMD and Pororo captions match the clips, they may not form complete storylines. The CVSV (Lu et al., 2023), YMS (Dogan et al., 2018), and SyMoN (Sun et al., 2022) datasets collect YouTube movie summaries, similar to M-SYMON, but CVSV and SyMoN lack human annotations.

Although there are many story understanding dataset, multilingual video story datasets are scarce. To our knowledge, the only dataset is Movie101v2 (Yue et al., 2024), which contains 46K Chinese video-caption pairs from 203 movies, where the Chinese text are translated to English using GPT-3.5-turbo. The dataset also lacks human-annotated video-text correspondence. In comparison, M-SYMON is large-scale and multilingual, containing movie summary videos in 7 languages and 101.5 hours of video with exact video-text correspondences.

2.3 Multilingual Story Understanding

There are several tasks related to multilingual story understanding, including event-causal inference (Lai et al., 2022), story question answering (Ateeq et al., 2023), story tag classification (Tikhonov et al., 2021), and story generation (Razumovskaia et al., 2024). However, most of them merely focus on text rather than video stories.

Methods for multilingual story understanding mainly fall into two categories: (1) translating training/test data to English (Ponti et al., 2020) or prompts to the target language (Lin et al., 2022), and (2) directly finetune on non-English text using a multilingual Pretrained Language Model (PLM) (Tikhonov et al., 2021) or a PLM for the target language (Lai et al., 2022). In this paper, we benchmark M-SYMON on multilingual story video-text alignment with both types of methods.

3 Task Formulation

Given a consecutive sequence of video clips $V = (v_1, v_2, ..., v_m)$ and a consecutive sequence of sentences $T = (t_1, t_2, ..t_n)$, the video-text alignment task aims to find the alignment between these two sequences by learning a function f that maps each input sentence t_i to its corresponding video clips:

$$\mathcal{P} = \{\dots, (t_i, v_{f(i)}), \dots\} \quad (1)$$

where $f(i)$ refers to the indexes of the video clip aligned with the sentence t_i. Note that $f(i)$ can be *None*, a single index, or an index sequence, denoting that there is no video clip, one video clip, or multiple video clips aligned with t_i, respectively.

In the task formulation, we make the simplifying assumption that multiple video clips can be aligned with one sentence, but a clip cannot be aligned with more than one sentence. The reason is that a video clip is typically very short (around 2.4 seconds) and clips outnumber sentences, so it is rare that one clip aligns with more than one sentence.

4 Dataset Construction

In this section, we describe the details of constructing the M-SYMON dataset.

4.1 Data Source and Statistics

For data collection, we first identify YouTube channels with movie recap videos in the target languages. Keywords such as "movie summary in <language>" and "movie recap in <language>" are used to search for the channels. We then download all videos and their accompanying subtitles from the identified channels. Videos without subtitles and that are not movie summaries are discarded.

This yields 13,166 videos or 2,136 video hours in 7 languages, including English, Chinese, Spanish, French, Portuguese, Hindi, and Russian. We

Languages	English	Chinese	Spanish	French	Portuguese	Hindi	Russian
Video Count	5,193	2,683	1,595	1,193	1,070	749	683
Video Hours	869	390	285	102	209	173	108
Vocabulary Size	40,116	6,269	50,050	33,967	40,676	20,896	64,827
Average Video Length (minutes)	9.5	8.7	10.7	5.1	11.7	13.9	9.5
Average number of Sentences in a Narration	131	78.8	82.6	138	134	126	151
Average number of Words in a Narration	1,717	3,234	1,923	2,226	2,737	2,656	1,623
Movie Count	2,440	1,236	1,217	877	811	462	542
Number of Annotated Videos	98	30	75	84	57	63	72
Annotated Video Hours	23.42	7.23	14.17	14.17	14.17	14.17	14.17

Table 1: The statistics of M-SYMON. For Chinese, we regard the number of unique characters as its vocabulary size.

list the statistics in Table 1. M-SYMON contains summaries for 5,960 movies and TV shows, of which 1,515 have more than one summary.

4.2 Human Annotation

We hire a professional team of annotators with rich translation experience from Flitto[2] to annotate exact video-text correspondence in M-SYMON. In total, we annotated 480 videos spanning 101.5 hours. The amount of annotated videos in each language are in the last two lines of Table 1.

Before human annotation, we automatically divided the video subtitle text into sentences. For each sentence, we ask the annotators to locate the start and end times of the video segment described by the sentence. If a sentence is not grounded in the video, it is marked as "unmatched". We removed the audio from the videos to eliminate shortcut features for alignment from the audio. Note that human-annotated video-text correspondence is different from temporal correspondence, because the narration in narrated movie summary videos is sometimes faster or slower than the video, and some narrator may add commentary not grounded in the video.

Annotation Quality. To evaluate annotation quality, we also employ a student who is familiar with the task to annotate a small validation set for each language. The annotation quality is evaluated as the IoU between the durations annotated by the student and the durations annotated by the annotators. If the IoU for a particular language falls below 60%, the annotators were asked to redo the annotations until the IoU reaches 60%. The final average IoU across 7 languages is 83.1%, indicating substantial agreement between the annotators. See Appendix B for annotation details.

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

4.3 Data Preprocessing

Multilingual Punctuation Restoration. We acquire text descriptions directly from YouTube subtitles. In some cases, the text descriptions are derived from automatic speech recognition tools and are unpunctuated. As punctuated texts are required for downstream tasks, we train a Transformer-based model to restore punctuations in all unpunctuated text. We do not use off-the-shelf punctuation restoration models because most models do not include all the languages in M-SYMON (Chordia, 2021; Frank and Böhme, 2021). See Appendix C for details on the punctuation restoration model.

Scene Segmentation. We divide the videos into clips using TransNet V2 (Souček and Lokoč, 2020) which detects hard camera cuts. Each clip, containing a continuous shot between two hard camera cuts, is roughly 2.4 seconds long.

Offensive Language Filtering. To the best of our abilities, we created a list of offensive terms and filtered out videos containing those phrases.

4.4 Dataset Split

We divide M-SYMON into two parts: (1) the human annotated portion: 480 videos from 7 languages manually annotated with exact video-text correspondence; (2) the weakly supervised portion: The entire M-SYMON dataset after removing all movies that appear in the annotated portion (see Appendix A.1 for details).

For the weakly supervised data, each video clip is matched to the text segment that spans its duration, producing a rough video-text correspondence regarded as the weakly supervised training set.

For the human annotated data, we further split it into a training set (20%), a validation set (20%), and a test set (60%), with the portions distributed evenly among 7 languages. Table 2 shows the number of clip-sentence pairs in each split.

Languages	English	Chinese	Spanish	French	Portuguese	Hindi	Russian
Weakly supervised training set	730,112	365,312	166,530	60,676	127,788	90,560	84,556
Supervised training set	6,649	1,860	5,989	5,799	4,205	7,152	5,789
Validation set	7,201	1,616	4,066	4,717	6,865	7,603	8,474
Test set	16,208	5,441	15,185	12,560	15,356	19,652	14,664

Table 2: The number of clip-sentence pairs in each data split.

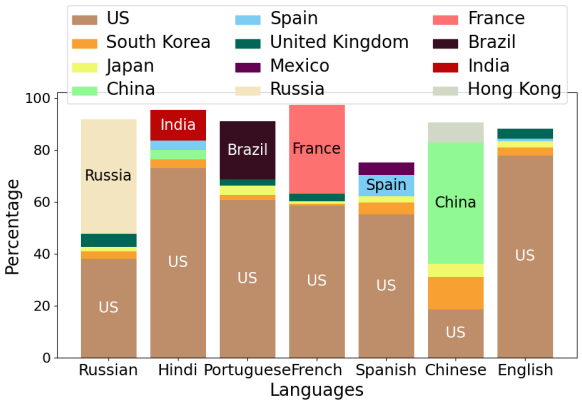

Figure 2: The top 5 countries of release for movies recapped in each language.

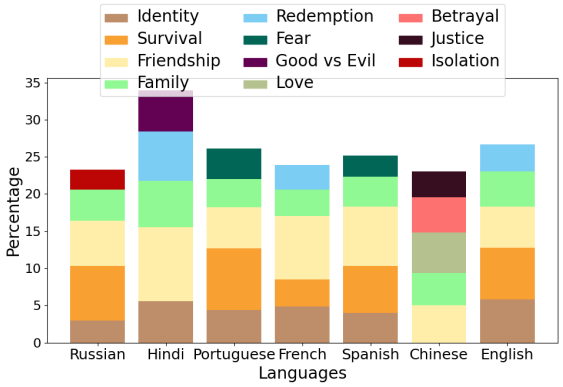

Figure 3: The top 5 themes for movies recapped in each language.

4.5 Dataset Analysis

We hypothesize that the choices of movies in each language may have cultural characteristics. Due to space limitations, we analyze the country of release and themes of the movies in this section. Additional statistics of the genre, year of release, language of release, and narrative structures are illustrated in Appendix A.2.

As most movie metadata are not available from the movie recap videos, we use ChatGPT (gpt-3.5-turbo-0125) to acquire the information. Specifically, we input the video title, the video-level description featured on YouTube, and first 5 sentences of the movie narration into ChatGPT and ask ChatGPT to output the movie title and metadata.

Figure 2 reveals interesting correlation between the language and the origin of the movie. The movie summaries in each language are typically about domestic movies (i.e., movies from countries where the language is widely used) or from the United States, reflecting the influence of Hollywood. For instance, 22.5% of Portuguese videos are about Brazilian movies. Interestingly, large percentage of videos in Chinese (46.9% from China Mainland, 7.6% from Hong Kong) and Russian (44.0%) describe domestic movies than other languages. In comparison, on average 20.4% of videos in other non-english languages describe domestic movies.

Figure 3 reveals variations in thematic preferences across languages. While themes like "Friendship" and "Family" appear universally popular, others exhibit uneven distributions. For instance, 2.7% (vs. 0.7% in other languages on average) of Russian movies have the theme "Isolation". The theme "Identity" accounts for only 1.8% Chinese videos (vs. an average of 4.6% in all other languages). These differences highlight the importance of representation of multiple cultures and languages in story understanding datasets.

5 Methodology

This section first details our base model for video-text alignment and then introduces its several variants with different multilingual training strategies.

5.1 Base Model

To achieve the sentence-level alignments between text and clip sequences, we adopt a three-stage framework widely used in recent studies (Dvornik et al., 2021; Zhang et al., 2023b). Specifically, our model first employs video and text encoders to obtain the video clip representation and the text representation as follows.

Video Encoder. Given a video with M clips, we

represent each clip by randomly selecting three frames from it and dividing each frame into $H \times W$ patches. We then utilize a trainable MLP projection layer to map these patches into video tokens, which are fed into the video encoder Swin Transformer (Liu et al., 2021) to obtain the visual representation of each frame. An average pooling layer is applied over the three frames of each video clip to obtain the video clip representation.

Text Encoder. Because our goal is to perform multilingual video-text alignment in this paper, we adopt a pre-trained cross-lingual cross-modal model named CCLM (Zeng et al., 2023) to initialize the parameters in Swin Transformer and the text encoder. Following CCLM, we employ a pre-trained multi-layer Transformer, i.e., XLM-R (Conneau et al., 2020) as the text encoder. We then regard the last hidden representation of the *[CLS]* token as the representation of each sentence. Formally, let $\boldsymbol{t}_i$ and $\boldsymbol{v}_i$ denote the encoded features for the i^{th} sentence and the i^{th} video clip.

For every clip-text pair, we randomly sample K hard negatives from the same video. With the sampled negative text features and video features, we finetune the encoders by minimizing the contrastive InfoNCELoss (Oord et al., 2018) below:

$$\mathcal{L}_{\text{InfoNCE}} = \frac{1}{N}\sum_{i=1}^{N}\left[-\log\left(\frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{t}_i/\tau)}{\sum_{j=1}^{K}\exp(\boldsymbol{v}_i^\top \boldsymbol{t}_j/\tau)}\right) - \log\left(\frac{\exp(\boldsymbol{v}_i^\top \boldsymbol{t}_i/\tau)}{\sum_{j=1}^{K}\exp(\boldsymbol{v}_j^\top \boldsymbol{t}_i/\tau)}\right)\right] \tag{2}$$

where N is the total number of training samples and K is the number of negative samples.

During inference, we first obtain the representation of each clip and sentence, and then calculate the cosine similarity between every clip-sentence pair. Finally, we find the global clip-sentence alignment from the similarity scores of each pair by resorting to a sequence alignment algorithm named Double Drop Dynamic Time Warp (Drop-DTW) (Dvornik et al., 2021), detailed in Appendix E.1.

5.2 Multilingual Finetuning Methods

With the base model, we benchmark the M-SYMON dataset by establishing 4 weakly supervised methods, which share the same model architecture but are trained on the entire weakly supervised training set of M-SYMON in Section 4.4 with different training strategies:

- Multilingual training (CCLM-multilingual). We combine the weakly supervised training sets of all 7 languages and train a unified multilingual model on the combined dataset.
- Individual training (CCLM-individual). Using weakly supervised training data, we finetune the base model on each language and obtain 7 language-specific models.
- Translational training (CCLM-translate). We first translate all the weakly supervised training data and the test data in M-SYMON to English with off-the-shelf translation model NLLB-200 (Costa-jussà et al., 2022). We then finetune the base model on the translated data to obtain a unified model.
- Two-stage training (CCLM-two-stage). On top of the trained CCLM-translate model, we further finetune on each language separately to obtain 7 language-specific models.

To show the usefulness of our manually annotated video-text alignment dataset, we further finetune two aforementioned methods on the human annotated training set in Section 4.4:

- CCLM-translate-supervision. We translate the non-English languages in the human annotated data to English and finetune the CCLM-translate model on the translated data.
- CCLM-two-stage-supervision. We finetune the CCLM-two-stage model on the human annotated data of each language separately.

6 Experiments

In this section, we conduct extensive experiments in intra-lingual, cross-lingual, and out-of-domain setups to show the usefulness of M-SYMON.

6.1 Experimental Setup

For CCLM-translate and CCLM-multilingual, we initialize their parameters from the pre-trained CCLM-base model (Zeng et al., 2023) and finetune them for 20 epochs with an initial learning rate of 4×10^{-5} and cosine learning rate decay. For CCLM-two-stage, we initialize its parameters from the CCLM-translate model and finetune it for 20 epochs with an initial learning rate of 4×10^{-6} and cosine learning rate decay. Random augmentation and weight decay of 0.2 are applied.

	English	Chinese	Spanish	French	Portuguese	Hindi	Russian	Average
				Weakly supervised				
CCLM-multilingual	9.3	16.3	11.8	8.8	9.9	5.8	9.3	10.2
CCLM-individual	26.6	29.3	17.4	16.6	16.3	10.1	14.6	18.7
CCLM-translate	22.2	20.1	16.0	17.5	14.1	11.2	13.8	16.4
CCLM-two-stage	***26.9***	***37.7***	***19.1***	***20.2***	***18.7***	**13.1**	***17.0***	***21.8***
				Supervised				
CCLM-translate-supervision	24.3	20.1	17.8	19.7	16.0	12.6	15.1	17.9
CCLM-two-stage-supervision	**27.7**	**38.9**	**19.8**	**21.0**	**19.5**	***12.9***	**17.5**	**22.5**

Table 3: Intra-lingual results based on F1 score. The "Average" is the mean of 7 languages. The highest number in each column is in **bold**, the second highest is in ***bold italic***.

6.2 Evaluation Metrics

Following Dogan et al. (2018), we use two evaluation metrics. Clip accuracy is defined as the temporal proportion of correctly aligned video segments. Sentence IoU (Jaccard, 1908) is defined as the intersection-over-union between the aligned video durations and the groundtruth durations. Due to space limitations, we mainly report the F1 score, i.e. the harmonic mean between Clip Accuracy and Sentence IoU in this section and defer detailed results to Appendix E.2.

6.3 Intra-Lingual Results

We report the intra-lingual results in Table 3. Here, each method is trained and evaluated on data in the same language.

First, we find that training the base model on all languages together performs significantly worse than training on each language separately. Specifically, CCLM-individual outperforms CCLM-multilingual by 4.3-17.3 percentage points. This suggests that language-specific features are important for multilingual video-text alignment. Second, the CCLM-translate baseline is on par with CCLM-individual and significantly outperforms CCLM-multilingual. Moreover, the two-stage baseline (i.e., CCLM-two-stage) leverages the benefits of pre-training on a large multilingual dataset and subsequent finetuning on a specific language dataset, which outperforms CCLM-individual and CCLM-translate by 3.1 and 5.4 percentage points on average. Finally, after finetuning with our manually annotated data, it yields consistent improvements on top of both CCLM-translate and CCLM-two-stage models across the 7 languages, and CCLM-two-stage-supervision achieves the best average performance among all the compared methods. These observations demonstrate the usefulness of the manually annotated video-text alignment data.

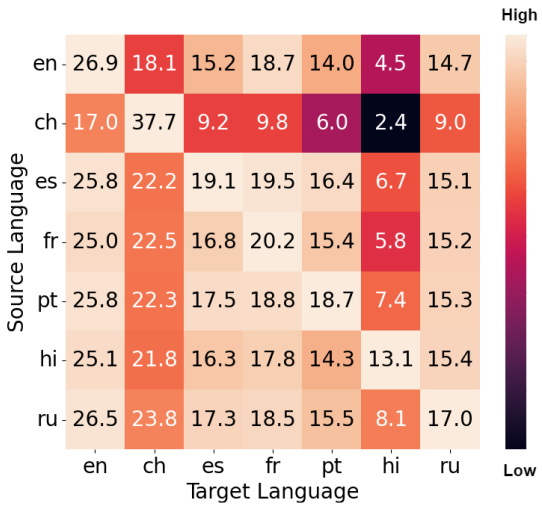

Figure 4: Cross-lingual transfer results of CCLM-two-stage based on F1 score. The language names are abbreviated as: English ="en", Chinese = "ch", Spanish = "es", French = "fr", Portuguese "pt", Hindi = "hi", Russian = "ru".

6.4 Cross-Lingual Transfer Results

We employ the CCLM-two-stage model to evaluate the cross-lingual performance of every source-target language pair, see Figure 4 for results.

First, the highest value of each column appears on the diagonal. This is intuitive because the model is trained and evaluated on the same language. Second, linguistically similar languages transfer well to each other. For example, related languages like Spanish and Portuguese, generally obtain good cross-lingual performance. This likely stems from shared vocabulary, sub-word tokens, and grammatical structures. Furthermore, Chinese does not transfer well to any other language, possibly due to differences in linguistic construction and movie distributions (See §4.5) . Lastly, we observe it is difficult to transfer from any language to Hindi, possibly because the CCLM pre-training data does not contain Hindi, which hurts representation learning.

	English	Chinese	Spanish	French	Portuguese	Hindi	Russian	Average
Without Timestamps of Text								
CCLM-two-stage-supervision	27.7	38.9	19.8	21.0	19.5	12.9	17.5	22.5
With Timestamps of Text								
Transcript baseline	34.4	42.6	19.1	26.2	**40.6**	20.1	**27.3**	30.1
CCLM-2S-supervision-time	**34.6**	**46.2**	**28.5**	**27.7**	39.8	**25.4**	26.4	**32.7**

Table 4: Intra-lingual video-text alignment results with transcript temporal information.

6.5 Effects of Timestamps of Narration Text

All previous models do not utilize the timestamps of the narration texts as input features or during inference. As such timestamps are only available after a video summary is made, models relying on timestamps as input features cannot handle general alignment tasks, such as aligning plot summaries with movies.

In this section, we explore the effectiveness of these timestamps during inference. Specifically, we use the CCLM-two-stage-supervised model to calculate video-text similarity. During DTW, we constrain video clips to match only sentences within a 1-second window of their timestamps. This model is named CCLM-2S-supervised-time. We also create a transcript baseline that relies solely on the timestamps; we simply align video clips to the sentence whose timestamp falls in their duration.

Table 4 summarizes the results. First, the transcript baseline achieves an average F1 score of only 30.1, indicating significant discrepancies between the temporal transcripts and our annotations. This demonstrates the noisy nature of YouTube transcripts and underscores the importance of precise annotations for accurate evaluation. Second, CCLM-2S-supervised-time outperforms CCLM-two-stage-supervised by 10.2 percentage points, showing that transcript temporal information improves alignment. Finally, CCLM-2S-supervised-time surpasses the transcript baseline by 2.6 percentage points, demonstrating that our model learns video-text correspondences beyond the transcripts.

6.6 Transfer to YMS

In this section, we extend our analysis to investigate if models trained on M-SYMON can generalize to other video-text alignment benchmarks, such as the out-of-domain English benchmark dataset YMS (Dogan et al., 2018).

As shown in Table 5, the large scale M-SYMON dataset significantly improves the performance of YMS. First, we can observe that finetuning the base model on the human-annotated training set of YMS (i.e., CCLM-individual (YMS)) is already on par with the state-of-the-art performance, demonstrating the effectiveness of our proposed base model. Second, compared to CCLM-individual (YMS), only training on the weakly supervised portion of M-SYMON improves performance by 13.7 and 12.7 percentage points on Clip Accuracy and Sentence IoU, respectively. This suggests the value of the large-scale weakly-supervised data in M-SYMON. Moreover, additional finetuning on the human-annotated data from M-SYMON, i.e., CCLM-two-stage-supervision (English), further boosts the performance by 2.4 and 3.0 percentage points on Clip Accuracy and Sentence IoU, respectively. These observations highlight the significance and generalizability of our human-annotated video-text correspondence dataset, indicating the potential of M-SYMON for advancing research in video story understanding.

	Clip Acc.	Sent IoU
NeuMatch (Dogan et al., 2018)	12.0	10.4
Wang et al. (2021a)	30.6	12.8
Sun et al. (2024)	23.2	18.4
CCLM-individual (YMS)	29.2	18.9
CCLM-multilingual	13.3	6.2
CCLM-individual (English)	35.8	25.0
CCLM-translate	25.2	16.7
CCLM-two-stage (English)	42.9	31.6
CCLM-two-stage-supervision (English)	**45.3**	**34.6**

Table 5: Out-of-domain results on the YMS dataset. CCLM-individual (YMS) is the base model finetuned on the human-annotated training set of YMS.

6.7 Discussion

The experiments in this paper demonstrate the value of M-SYMON as a multilingual and multimodal story understanding dataset. First, M-SYMON is the first dataset that offers large-scale human annotations for multilingual story video-text alignment, and such annotations deliver pratical benefits. As shown in §6.5, the timestamps

in the YouTube transcripts are noisy, and human annotations are necessary for accurate benchmarking. Additionally, the human-annotated training set provides valuable supervision signals. In §6.3, we achieve a notable 3.2% relative improvement from human annotation data that amounts to only 2.2% of the training data. Moreover, finetuning on supervised data from M-SYMON lead significant improvement on the out-of-domain YMS dataset in §6.6. Finally, M-SYMON provides fertile ground for linguistic and cultural studies of story elements. For example, cross-lingual transfer experiments (§6.4) lends support to existing linguistic theories on positive and negative transfer in language learning (Eronen et al., 2023; Odlin, 1989) as linguistically similar languages transfer well to each other. The analysis of movie themes for different languages (Figure 3) demonstrates the cultural relevance of M-SYMON.

7 Conclusion

In this paper, we introduced M-SYMON, a large-scale multilingual video story dataset. It contains 13,166 movie summary videos from 7 languages, featuring 480 videos with human-annotated video-text correspondence. We established several multilingual multimodal baselines to benchmark M-SYMON. Experimental results show the value of M-SYMON for video story understanding. Its size, multilingual nature, and rich alignment annotations make M-SYMON a valuable contribution.

Acknowledgments

We gratefully acknowledge the support by the Nanyang Associate Professorship and the National Research Foundation Fellowship (NRF-NRFF13-2021-0006), Singapore. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of the funding agencies.

Limitations

The preprocessing of M-SYMON involves several automatic techniques that may introduce noise. First, some text descriptions are generated using Google automatic speech recognition. Automatic punctuation restoration is then applied to the text. For the translation baselines, the text is further translated into English. This pipeline may propagate errors. However, due to the size and complexity of M-SYMON, manual processing is impractical. We acknowledge the potential for improvement in preprocessing steps as better techniques become available.

Additionally, human annotation of video-text correspondence in story videos can be ambiguous. Annotators received a list of instructions for ambiguous cases. For example, when text describes a character's emotional state, annotators are instructed to match only when the emotion is evident from the character's expression or action. In cases outside the provided guidelines, annotators used their best judgment. Despite the inherent ambiguity in story content annotation, manual inspection of the annotated data indicates good quality, as shown in Table 6.

Ethics Statement

M-SYMON is constructed with YouTube videos under the Standard YouTube License. For the videos, we release a list of YouTube URL the users can use to download the video from YouTube, as is the standard practice (Bain et al., 2020; Miech et al., 2019; Caba Heilbron et al., 2015). A fair compensation amount for the annotators was determined by the annotation company based on the difficulty level and time needed to annotate a minute of video in each language. On average, we compensate the annotator 2.44 USD for annotating a minute of video. The exact per-minute compensation amount for each language is shown in Table 7 in the Appendix.

In this paper, we collect user-uploaded videos from YouTube, which are summaries of mostly western movies and TV shows. We recognize that movies and TV shows are fictional in nature, and often prioritize dramatic events over faithful representation of real-life scenarios. In addition, the videos may reflect particular bias of the creators of the movie and TV shows or the creators of the summary videos, as well as bias from particular cultures or the time periods of production.

For these reasons, we urge researchers to take caution when attempting to learn social norms from such videos. For example, events of bank robberies may be over-represented in these videos, and a machine learning model may inadvertently infer that robbing a bank is part of the social norm. In addition, the model may incorrectly learn from disproportional association of certain groups of people with certain social status, occupations, and other cultural constructs. The dataset is intended for fun-

damental research and not real-world deployment.

References

Berkeley R Andrus, Yeganeh Nasiri, Shilong Cui, Benjamin Cullen, and Nancy Fulda. 2022. Enhanced story comprehension for large language models through dynamic document-based knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10436–10444.

Mohammad A Ateeq, Sabrina Tiun, Hamed Abdelhaq, and Nawras Rahhal. 2023. Arabic narrative question answering (qa) using transformer models. *IEEE Access*.

Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. 2020. Condensed movies: Story based retrieval with contextual embeddings. In *ACCV*.

Yogesh Balaji, Martin Renqiang Min, Bing Bai, Rama Chellappa, and Hans Peter Graf. 2019. Conditional gan with discriminative filter generation for text-to-video synthesis. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 1995–2001.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III au2, and Chris Dyer. 2016. Modeling dynamic relationships between characters in literary novels. In *AAAI*.

Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. 2021. Dramaqa: Character-centered video story understanding with hierarchical qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1166–1174.

Varnith Chordia. 2021. Punktuator: A multilingual punctuation restoration system for spoken and written text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 312–320.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Timothee Cour, Chris Jordan, Eleni Miltsakaki, and Ben Taskar. 2008. Movie/script: Alignment and parsing of video and text transcription. In *Computer Vision–ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV 10*, pages 158–171. Springer.

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

Yijiang Dong, Lara Martin, and Chris Callison-Burch. 2023. Corrpus: Code-based structured prompting for neurosymbolic story understanding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13152–13168.

Li Du, Xiao Ding, Ting Liu, and Bing Qin. 2021. Learning event graph knowledge for abductive reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5181–5190.

Mikita Dvornik, Isma Hadji, Konstantinos G Derpanis, Animesh Garg, and Allan Jepson. 2021. Drop-dtw: Aligning common signal between sequences while dropping outliers. *Advances in Neural Information Processing Systems*, 34:13782–13793.

David Elson, Nicholas Dames, and Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 138–147.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Juuso Eronen, Michal Ptaszynski, and Fumito Masui. 2023. Zero-shot cross-lingual transfer language selection using linguistic similarity. *Information Processing & Management*, 60(3):103250.

Oliver Guhr1 Anne-Kathrin Schumann Frank and Bahrmann1 Hans-Joachim Böhme. 2021. Fullstop: Multilingual deep models for punctuation prediction.

Xu Gu, Yuchong Sun, Feiyue Ni, Shizhe Chen, Xihua Wang, Ruihua Song, Boyuan Li, and Xiang Cao. 2023. Tevis: Translating text synopses to video storyboards. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4968–4979.

Mandy Guo, Zihang Dai, Denny Vrandečić, and Rami Al-Rfou. 2020. Wiki-40B: Multilingual language model dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2440–2452, Marseille, France. European Language Resources Association.

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

Tengda Han, Weidi Xie, and Andrew Zisserman. 2022. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916.

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

Qiuyuan Huang, Zhe Gan, Asli Celikyilmaz, Dapeng Wu, Jianfeng Wang, and Xiaodong He. 2019. Hierarchically structured reinforcement learning for topically coherent visual story generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8465–8472.

Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270.

Evgeny Kim and Roman Klinger. 2019. Frowning frodo, wincing leia, and a seriously great friendship: Learning to classify emotional relationships of fictional characters. In *NAACL*.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. In *IJCAI*, pages 2016–2022.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Meci: A multilingual dataset for event causality identification. In *Proceedings of the 29th international conference on computational linguistics*, pages 2346–2356.

Zeqian Li, Qirui Chen, Tengda Han, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. A strong baseline for temporal video-text alignment. *arXiv preprint arXiv:2312.14055*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Yue Liu, Xin Wang, Yitian Yuan, and Wenwu Zhu. 2019. Cross-modal dual learning for sentence-to-video generation. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1239–1247.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.

Yu Lu, Feiyue Ni, Haofan Wang, Xiaofeng Guo, Linchao Zhu, Zongxin Yang, Ruihua Song, Lele Cheng, and Yi Yang. 2023. Show me a video: A large-scale narrated video dataset for coherent story illustration. *IEEE Transactions on Multimedia*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Terence Odlin. 1989. *Language transfer*, volume 27. Cambridge University Press Cambridge.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Evgeniia Razumovskaia, Joshua Maynez, Annie Louis, Mirella Lapata, and Shashi Narayan. 2024. Little red riding hood goes around the globe: Crosslingual story planning and generation with large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10616–10631, Torino, Italia. ELRA and ICCL.

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

Yisi Sang, Xiangyang Mou, Mo Yu, Shunyu Yao, Jing Li, and Jeffrey Stanton. 2022. Tvshowguess: Character comprehension in stories as speaker guessing. In *NAACL*.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 2022. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5026–5035.

Tomáš Souček and Jakub Lokoč. 2020. Transnet v2: An effective deep network architecture for fast shot transition detection.

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

Yidan Sun, Qin Chao, and Boyang Li. 2024. Event causality is key to computational story understanding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3493–3511.

Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. 2015. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1827–1835.

Alexey Tikhonov, Igor Samenko, and Ivan Yamshchikov. 2021. StoryDB: Broad multi-language narrative dataset. In *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*, pages 32–39, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jianan Wang, Boyang Li, Xiangyu Fan, Jing Lin, and Yanwei Fu. 2021a. Data-efficient alignment of multimodal sequences by aligning gradient updates and internal feature distributions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 665–675.

Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021b. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5079–5088.

Chao-Yuan Wu and Philipp Krahenbuhl. 2021. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894.

Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Li, Nora Bradford, Branda Sun, Tran Hoang, Yisi Sang, Yufang Hou, Xiaojuan Ma, Diyi Yang, Nanyun Peng, Zhou Yu, and Mark Warschauer. 2022. Fantastic questions and where to find them: FairytaleQA – an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460, Dublin, Ireland. Association for Computational Linguistics.

Zihao Yue, Yepeng Zhang, Ziheng Wang, and Qin Jin. 2024. Movie101v2: Improved movie narration benchmark. *arXiv preprint arXiv:2404.13370*.

Yan Zeng, Wangchunshu Zhou, Ao Luo, Ziming Cheng, and Xinsong Zhang. 2023. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. pages 5731–5746.

Heng Zhang, Daqing Liu, Zezhong Lv, Bing Su, and Dacheng Tao. 2023a. Exploring temporal concurrency for video-language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15568–15578.

Jiahao Zhang, Anoop Cherian, Yanbin Liu, Yizhak Ben-Shabat, Cristian Rodriguez, and Stephen Gould. 2023b. Aligning step-by-step instructional diagrams to video demonstrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2483–2492.

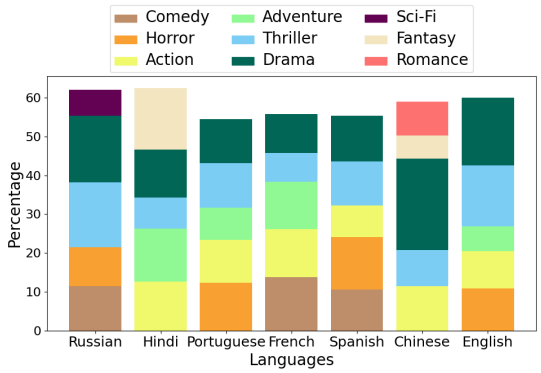

Figure 5: The top 5 genres for movies recapped in each language

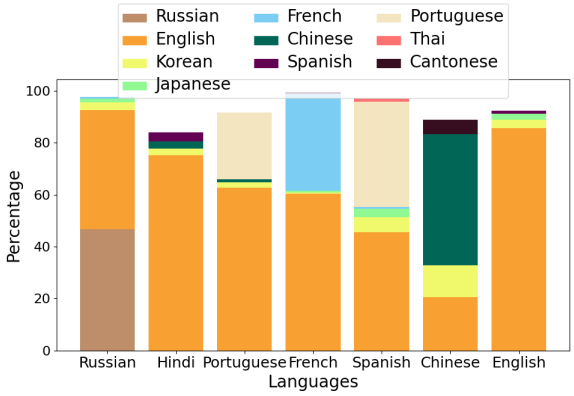

Figure 6: Top 5 languages of release for movies recapped in each language

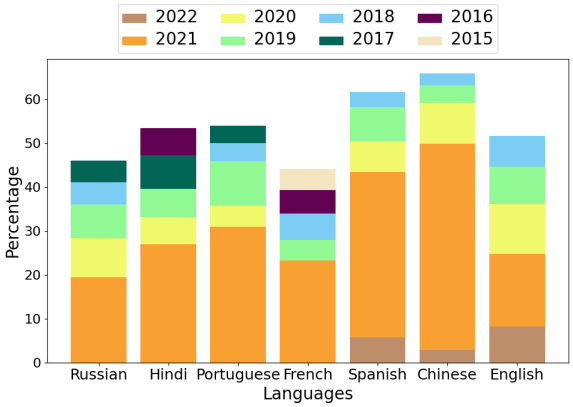

Figure 7: Top 5 years of release for movies recapped in each language

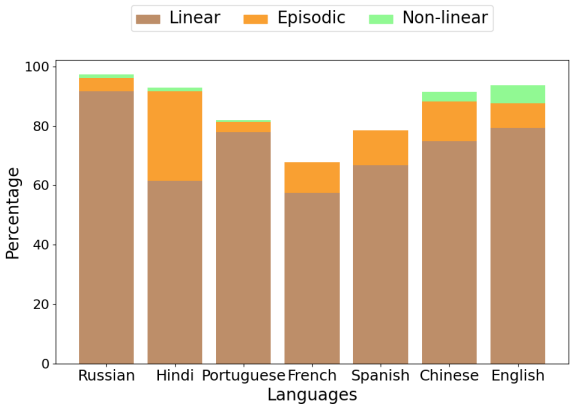

Figure 8: Narrative style of movies recapped in each language

A Additional Information on M-SYMON

A.1 Additional Preprocessing

Removing Overlap Between Train and Test Sets. In movie recap videos, videos summarizing the same movie may have similar visual and textual content. Thus, we remove all movies that appear in the test sets from the training sets. Specifically, we use ChatGPT to identify the corresponding movie for each video and translate the movie name to English. We then remove the overlapped videos based on the English movie names.

A.2 Additional Analysis

As shown in Figure 5, creators using different languages prefer similar genre. Although interestingly, "Fantasy" films are featured more in Hindi and Chinese than other language, perhaps due to the significant role mythology plays in Eastern cultures.

In Figure 6, we list the language of release for movies recapped in each language. This trend largely follows the same pattern as the country of release likely because most countries produce movies in their native language.

For each language, we list the top 5 years the movies were released (Figure 7). As our dataset collection ends around February 2022, M-SYMON largely contains movies released from 2017 to 2021.

In Figure 8, we show the narrative styles of movies recapped in each language. The most common narrative styles are linear and episodic.

B Annotation Details

Characteristics Of Annotators. We employ a team of annotators with professional qualifications in translation for annotation. The team is from Korea. We discussed the compensation amount with the annotation team and set an adequate amount.

Annotation Instruction. We gave the annotators an instruction, shown in Figure 9, and a set of annotation guidelines for ambiguous situations.

```
You will see a 5-15 minute video, which is
a summary of a movie or a TV show. You will
be given a list of sentences that describe
the content of the video. Please align each
sentence to the video content by writing
down the start time and end time of the
video segment that matches the sentence.
If the sentence does not match any video
content, please mark it as “Unmatched”.
```

Figure 9: Human annotation instruction.

Annotation Procedure. We divide the data randomly into 4 equal batches, each batch contains a quarter of data from each language. After the annotators annotate each batch, we randomly select 2 videos from each language for validation and employ a student familiar with the task to annotate them. Then, we calculate the IoU between the student's annotations and the annotations provided by the annotators. Specifically, for each sentence we calculate the IoU between the duration annotated by the student and the duration annotated by the annotators. If the IoU on a particular language falls below 60%, the annotators were asked to redo the annotations for that language in that batch. The average IoU of each language is shown in Table 6

C Multilingual Punctuation Restoration

Task. We treat punctuation restoration as a token classification task, where the model predicts the appropriate punctuation, if any, to follow each token in the sequence. To simplify the task, we limit the predictions to three punctuations: period, comma, and question mark. Other punctuations are either removed (in the case of #,@," and ') or replaced with periods and commas (in the case of !, ; and :).

Dataset. For training, we use the Wiki-40B (Guo et al., 2020) dataset containing clean Wikipedia articles in 40+ languages. For Hindi, we further supplement the training set with data from Wikipedia Hindi [3].

To evaluate our punctuation restoration model on narrative content, we build an evaluation set from Wikipedia plot summaries. Specifically, we first identify Wikipedia articles of popular movies. Then, we extract the "Plot Summary" portion of the articles. Finally, we remove any overlap between training and test data. We acquire the evaluation sets for French, Spanish, Portuguese, Chinese, and Russian following the above procedure. However, for Hindi, we can not find enough Wikipedia plot summaries. Therefore, we use the wiki-40B validation set for evaluation which may inflate performance on Hindi.

Model. Following (Frank and Böhme, 2021), we finetune on the XLM-RoBERTa-large (Conneau et al., 2019) model.

Evaluation. We evaluate the model performance with the F1 score.

Baselines. We use the Full-Stop (Frank and Böhme, 2021) model as the baseline [4]. We use the same backbone as Full-Stop, the difference is that they train on the Europarl (Koehn, 2005) dataset while we train on Wiki-40B. The Europarl dataset contains transcripts of political talks in European languages and is commonly used in training multilingual punctuation restoration models.

Results. Table 8 shows the multilingual punctuation model performance. Our model achieves good performance on all languages. On European languages, our model outperforms Full-Stop by 6-14% (see Table 9). This demonstrates that our model is suitable for punctuation restoration in narratives. The performance on Hindi is superior to other languages, this is likely because Hindi is evaluated on in-domain data from Wiki-40B.

In Table 14, we show the Clip Accuracy and Sentence IoU for the intra-lingual experiments.

In Tables 15 and 16 we show the Clip Accuracy and Sentence IoU scores for cross-lingual transfer, corresponding to Figure 4

D Video-Text Retrieval

Task. Given a set of video clips $\mathcal{V} = \{V_1, V_2, \ldots, V_M\}$ and a set of textual sentences $\mathcal{T} = \{t_1, t_2, \ldots, t_N\}$, both from all videos in the test set, the goal is to retrieve the most relevant video or video clips corresponding to each sentence.

Evaluation. We evaluate with Retrieval@1 (R@1), Retrieval@5 (R@5), Retrieval@10 (R@5) and Mean Rank.

[3] https://www.tensorflow.org/datasets/catalog/wikipedia

[4] The original Full-Stop model is trained on English, German, French, and Italian. We follow their instruction to train on Spanish, French, and Portuguese

	English	Chinese	Spanish	French	Portuguese	Hindi	Russian
IoU	83.2%	78.8%	84.4%	81.3%	88.9%	84.4%	80.8%

Table 6: IoU between annotator annotations and our annotations.

	English	Chinese	Spanish	French	Portuguese	Hindi	Russian
Annotation Cost	1.60	1.44	2.83	3.78	2.83	1.79	2.83

Table 7: Annotator compensation amount for annotating a minute of video, in USD.

	0	.	,	?	Total
French	0.98	0.86	0.71	0.56	0.78
Spanish	0.97	0.77	0.67	0.53	0.74
Portuguese	0.98	0.83	0.72	0.50	0.76
Russian	0.96	0.73	0.77	0.51	0.74
Chinese	0.99	0.82	0.83	0.61	0.81
Hindi	0.99	0.94	0.88	0.77	0.90
Average	0.98	0.82	0.76	0.58	0.79

Table 8: Result of multi-lingual punctuation.

	0	.	,	?	Total
Wiki-40B (Ours)					
French	0.98	0.86	0.71	0.56	0.78
Spanish	0.97	0.77	0.67	0.53	0.74
Portuguese	0.98	0.83	0.72	0.50	0.76
Europarl					
French	0.97	0.80	0.64	0.47	0.72
Spanish	0.97	0.70	0.59	0.15	0.60
Portuguese	0.97	0.75	0.63	0.37	0.68

Table 9: Comparison between training on Wiki-40B and training on Europarl. Europarl only contains European languages.

Method. For video-text retrieval, we evaluate on the CCLM models trained on M-SYMON. Specifically, we first acquire the video and text embeddings from the video and text encoders. Then we calculate the video text similarity as the cosine similarity between video and text embeddings. Finally, for each text we retrieve the video clip with the highest similarity.

Results. In Tables 10,11.12,13 we show R@1, R@5, R@10 and MR scores, respectively.

The text-to-video retrieval results align closely with our alignment results. CCLM-two-stage outperforms CCLM-translate by 2.2-6.5% on Retrieval @ 1,5,10, demonstrating the advantage of multi-lingual data. CCLM-two-stage-supervision outperforms CCLM-two-stage by 0.2-1.1% on Retrieval @ 1,5,10, highlighting the utility of our human annotations. However, supervised finetuning does not improve Spanish performance. We believe this may be due to the greater distribution difference between weakly supervised correspondence, which rely on timestamps, and human-annotated correspondence. Based on our statistics, only 34.0% timestamp-derived correspondence in Spanish videos match human annotations, while the average across all languages is 44.9%. Given this large distribution gap, more supervised data may be required to bridge the gap.

E Video-Text Alignment

E.1 Drop-DTW

DTW uses dynamic programming to find the optimal alignment between two sequences based on distance (or similarity), the final alignment corresponds to the shortest distance or highest similarity. In the traditional DTW algorithm, each item in one sequence must match with an item in the other sequence. However, in story videos, some text are not grounded in the video and vice-versa. Therefore, we use the Drop-DTW (Dvornik et al., 2021) algorithm to facilitate dropping certain clips and sentences.

In traditional DTW, to align a sequence of video clips $V = (v_1, \ldots, v_N)$ and a sequence of sentences $T = (t_1, \ldots, t_M)$, we first assume that v_1 is aligned to t_1. Thus, the cost of matching v_1 and t_1 is $c(1,1) = 0$, and the cost of match v_1 with $t_j (j \neq 1)$ is $c(1,j) = \infty$, and vice versa. Then, the minimal cost of aligning $(v_1, \ldots, v_i)$ with $(t_1, \ldots, t_j)$, can be calculated as:

$$\begin{aligned} c(i,j) = \min(&c(i-1,j) + d(i,j), \\ &c(i,j-1) + d(i,j), \\ &c(i-1,j-1) + d(i,j)) \end{aligned} \tag{3}$$

Model	English	Chinese	Spanish	French	Portuguese	Hindi	Russian	Average
Weakly supervised								
CCLM-multilingual	0.5	7.9	1.6	1.1	1.2	1.0	1.8	2.2
CCLM-individual	3.3	6.5	3.3	1.8	2.9	2.0	3.7	3.4
CCLM-translate	1.7	4.7	2.4	1.9	1.9	2.1	3.1	2.5
CCLM-two-stage	3.6	11.5	3.6	3.0	3.2	2.4	5.3	4.7
Supervised								
CCLM-translate-supervision	2.1	3.1	2.6	2.6	1.8	2.2	3.8	2.6
CCLM-two-stage supervision	3.7	13.8	3.6	3.3	3.0	2.9	4.2	4.9

Table 10: Text-to-Video retrieval results based on Retrieval@1.

Model	English	Chinese	Spanish	French	Portuguese	Hindi	Russian	Average
Weakly supervised								
CCLM-multilingual	1.6	34.5	4.9	6.2	4.3	3.9	1.6	8.1
CCLM-individual	10.6	17.4	10.3	5.3	7.9	6.6	10.6	9.8
CCLM-translate	6.0	12.9	7.5	6.6	6.1	8.5	6.0	7.7
CCLM-two-stage	10.7	30.3	12.2	8.8	10.0	8.3	10.7	13.0
Supervised								
CCLM-translate-supervision	7.1	12.2	8.3	7.6	6.2	7.5	7.1	8.0
CCLM-two-stage supervision	10.7	34.2	12.0	10.1	10.6	8.3	10.7	13.8

Table 11: Text-to-Video retrieval results based on Retrieval@5.

where $d(i, j)$ denotes the distance between v_i and t_j. Since we have the cosine similarity between each video-text pairs, the distance can be calculated as $d(i, j) = 1 - s(i, j)$, where $s(i, j)$ is the cosine similarity between v_i and t_j

In Drop-DTW, we further define two hyperparameters d_v and d_t as the cost of dropping the video and the text respectively. At each time stamp, we calculated the minimal cost as the lowest cost between: (1) matching v_i and t_j and adding $d(i, j)$ to the total cost; (2) dropping v_i and adding d_v to the total cost; (3) dropping t_j and adding d_t to the cost; (4) dropping both v_i and t_j and adding both d_v and d_t to the cost.

E.2 Addition results

In Table 14 we show the Clip Accuracy and Sentence IoU scores corresponding to Table 3.

E.3 Licensing Information

The videos we acquire from YouTube are under the Standard YouTube license. Note that in M-SYMON, the videos are released in the form of YouTube URL, and researchers can download the videos directly from YouTube.

For multilingual punctuation restoration we train with data from Wiki-40B (Guo et al., 2020) licensed under the Apache 2.0 License. For scene segmentation, we use TransNet-V2 (Souček and Lokoč, 2020) released under the MIT license. ChatGPT, which we use for extracting movie names and metadata, is under the GNU Affero General Public License Version 3.

For video text alignment, the CCLM (Zeng et al., 2023) is under the BSD-3-Clause license and the CLIP4Clip is under the MIT License. YMS (Dogan et al., 2018), which we use as an out-of-domain evaluation benchmark, is from `https://github.com/RubbyJ/Data-efficient-Alignment`.

The usage of all model and data within the paper are in line with their intended uses.

Model	English	Chinese	Spanish	French	Portuguese	Hindi	Russian	Average
Weakly-supervised								
CCLM-multilingual	2.5	38.4	8.2	3.6	7.1	7.1	9.5	10.9
CCLM-individual	16.1	27.3	16.3	9.1	12.7	9.6	20.3	15.9
CCLM-translate	9.6	21.1	13.2	11.1	10.2	13.6	15.6	13.5
CCLM-two-stage	16.0	40.8	19.2	14.3	15.9	12.2	21.4	20.0
Supervised								
CCLM-translate-supervision	12.1	19.9	13.0	12.0	10.0	13.1	16.7	13.8
CCLM-two-stage supervision	16.6	45.1	18.3	15.7	16.5	13.9	21.9	21.1

Table 12: Text-to-Video retrieval results based on Retrieval@10.

Model	English	Chinese	Spanish	French	Portuguese	Hindi	Russian	Average
Weakly-Supervised								
CCLM-multilingual	107	65	422	805	518	409	367	524
CCLM-individual	315	74	220	444	311	289	163	259
CCLM-translate	525	105	292	402	344	259	176	300
CCLM-two-stage	333	55	190	337	255	254	136	223
Supervised								
CCLM-translate-supervision	427	108	237	330	334	249	167	265
CCLM-two-stage supervision	326	47	182	282	254	246	129	209

Table 13: Text-to-Video retrieval results based on Mean Rank.

	English		Chinese		Spanish		French		Portuguese		Hindi		Russian	
	Clip.	Sent.	Clip.	Sent.	Clip.	Sent.	Clip.	Sent.	Clip.	Sent.	Clip.	Sent.	Clip.	Sent.
Weakly supervised														
CCLM-multi	13.4	7.3	23.5	12.8	19.5	8.9	17.6	8.8	17.4	7.2	11.8	3.8	15.1	7.0
CCLM-individual	33.2	22.3	36.6	24.6	25.7	13.6	24.2	11.2	24.8	12.4	17.7	7.3	21.9	11.2
CCLM-translate	28.4	18.4	27.7	15.9	24.0	12.5	24.3	13.8	21.9	10.7	18.6	8.2	20.1	10.8
CCLM-two-stage	**33.7**	**22.6**	**43.7**	**33.3**	**27.2**	**15.2**	**27.8**	**16.3**	**27.0**	**14.6**	**20.9**	**9.9**	**24.5**	**13.3**
Supervised														
CCLM-translate-s	30.9	20.0	27.5	20.1	26.5	14.0	26.9	15.6	24.4	12.3	20.9	9.3	21.9	11.8
CCLM-two-stage-s	**34.6**	**23.5**	**45.2**	**34.4**	**27.4**	**15.9**	**28.3**	**16.9**	**27.7**	**15.4**	20.5	9.6	**25.1**	**13.7**

Table 14: Intra-lingual results based on Clip Accuracy and Sentence IoU. Due to space limitation, we write Clip Accuracy as "Clip.", Sentence IoU as "Sent." , "-supervised" as "-s", and "multilingual" as "multi".

	English	Chinese	Spanish	French	Portuguese	Hindi	Russian
CCLM-translated	28.4	27.7	24.0	24.3	21.9	18.6	20.1
CCLM-two-stage (English)	**33.7**	25.4	23.3	25.9	21.1	15.1	21.4
CCLM-two-stage (Chinese)	22.4	**43.7**	16.7	14.8	11.8	12.0	15.3
CCLM-two-stage (Spanish)	32.5	29.9	**27.2**	26.6	24.2	14.8	22.0
CCLM-two-stage (French)	31.7	30.9	25.1	**27.8**	23.1	15.0	22.5
CCLM-two-stage (Portuguese)	32.6	30.2	25.4	27.1	**27.0**	16.6	22.2
CCLM-two-stage (Hindi)	31.8	29.8	25.2	26.1	21.7	**20.9**	22.7
CCLM-two-stage (Russian)	33.2	31.7	25.6	26.5	23.5	16.6	**24.5**

Table 15: Clip Accuracy scores for cross-lingual transfer.

	English	Chinese	Spanish	French	Portuguese	Hindi	Russian
CCLM-translated	18.4	15.9	12.5	13.8	10.7	8.2	10.8
CCLM-two-stage (English)	**22.6**	14.2	11.7	14.8	10.8	2.9	11.5
CCLM-two-stage (Chinese)	13.8	**33.3**	6.7	7.6	4.2	1.4	6.6
CCLM-two-stage (Spanish)	21.5	18.0	**15.2**	15.6	12.8	4.6	11.7
CCLM-two-stage (French)	20.8	17.8	13.1	**16.3**	11.8	3.9	15.2
CCLM-two-stage (Portuguese)	21.5	17.9	13.7	15.7	**14.6**	5.2	11.8
CCLM-two-stage (Hindi)	20.8	17.4	12.5	14.8	11.0	**9.9**	11.8
CCLM-two-stage (Russian)	22.1	19.2	13.6	15.1	12.0	5.7	**13.3**

Table 16: Sentence IoU scores for cross-lingual transfer.