Translating Text Synopses to Video Storyboards

Xu Gu[1], Yuchong Sun[1], Feiyue Ni[1], Shizhe Chen[2], Ruihua Song[1], Boyuan Li[1], Xiang Cao[3]
[1]Renmin University of China, Beijing, China,
[2]Inria, École normale supérieure, CNRS, PSL Research University,
[3]Bilibili Corporation, Shanghai, China
https://ruc-aimind.github.io/projects/tevis.html

Abstract

*A storyboard is a roadmap for video creation which consists of shot-by-shot images to visualize key plots in a text synopsis. Creating video storyboards however remains challenging which not only requires association between high-level texts and images, but also demands for long-term reasoning to make transitions smooth across shots. In this paper, we propose a new task called **Te**xt synopsis to **Vi**deo **S**toryboard (**TeViS**) which aims to retrieve an ordered sequence of images to visualize the text synopsis. We construct a **MovieNet-TeViS** benchmark based on the public MovieNet dataset [15]. It contains 10K text synopses each paired with keyframes that are manually selected from corresponding movies by considering both relevance and cinematic coherence. We also present an encoder-decoder baseline for the task. The model uses a pretrained vision-and-language model to improve high-level text-image matching. To improve coherence in long-term shots, we further propose to pre-train the decoder on large-scale movie frames without text. Experimental results demonstrate that our proposed model significantly outperforms other models to create text-relevant and coherent storyboards. Nevertheless, there is still a large gap compared to human performance suggesting room for promising future work.* [1]

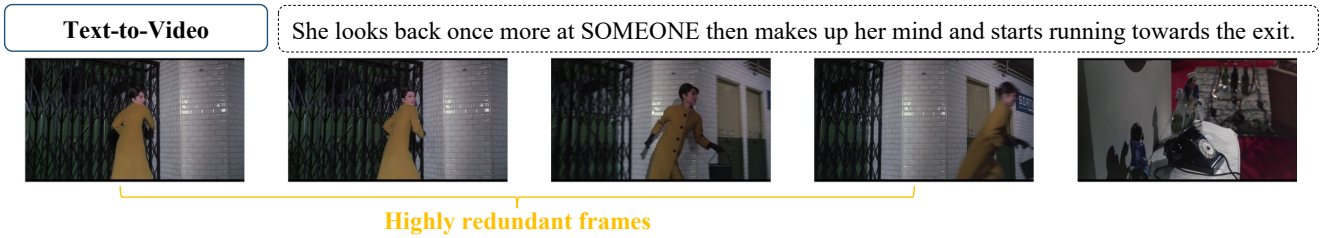
1. Introduction

With the prevalence of video sharing platforms, more and more video creators are emerging with enthusiasm to create videos using their own text synopses. An initial and critical step in professional video creation is to translate a text synopsis into a video storyboard, which is a sequence of shot-by-shot images to visualize key plots in a screenplay. Creating a high-quality video storyboard is however challenging for amateurs. It not only requires one to put relevant scenes, characters and actions in the video, but also demands for cinematic organizations of keyframes such as coherent transitions across shots *etc*. Hence, there are high application needs in assisting amateurs to create more professional video storyboards from their text synopses.

Although existing works have made great progress in text-to-image retrieval [8,10,18,19,25,45], text-to-video retrieval [2,22,23,44] and even text-to-video generation [14,35,42], they are limited in creating storyboards from texts. The text-to-image retrieval works can only produce static images without considering the dynamics of shots in the video. Text-to-video works are able to retrieve or generate videos. Yet, most of them focus on short-term video clips with only a few seconds as shown in Fig. 1a. The images in these videos are highly redundant and cannot satisfy the requirement of a video storyboard for coherent keyframes [1,32]. The visual storytelling works [7,16,30] are proposed to visualize text with a sequence of images, but they care more about the text-image relevancy while omitting long-term reasoning to make transition smooth across keyframes (see Fig. 1b). Moreover, the query texts in existing works are visually concrete and descriptive, making the models less generalizable to more abstract and high-level text synopses such as the synopsis in Fig. 1c.

In order to reduce the gap between existing tasks and realistic needs for storyboard creation, in this work, we propose a new task called **Te**xt synopsis to **Vi**deo **S**toryboard (**TeViS**). In the TeViS task, we aim to retrieve an ordered sequence of images from large-scale movie database as video storyboard to visualize an input text synopsis. For this purpose, we collect the **MovieNet-TeViS** benchmark based on the public MovieNet dataset [15]. The MovieNet dataset contains high-level text synopses for movies and a coarse-grained alignment between movie segments and text synopses paragraphs. We ask annotators to split paragraphs into semantically compact sentences and select a minimum set of keyframes from its aligned movie segment for each text synopsis sentence. Annotators should consider both relevancy to the text and cinematic coherency across frames for keyframe selection. Finally, we obtain 10K text syn-

[1]We will release our dataset for future research.



(a) Example from the LSMDC dataset [32] with detailed text descriptions and short-term video clips.

(b) Example from the VIST dataset [16] with story descriptions and incoherent image sequences.

(c) Our constructed MovieNet-TeViS dataset for video storyboard creation from text synopsis.

Figure 1. Comparison of our proposed Text Synopsis to Video Storyboard task and existing tasks.

opses and each paired with 4.6 keyframes on average.

There are two unique challenges posed in our TeViS task. First, the text synopses are diverse covering a wide range of topics and some of them are also high-level and abstract, *e.g.*, 2.76 concreteness score on average (vs. 2.99 in other video-text datasets such as LSMDC [32] and MAD [36]). Therefore, it is much more difficult to visualize texts with relevant images. Second, our text synopses correspond to video storyboards with much longer time range, *e.g.*, 64 seconds on average (vs. 4 seconds in LSMDC or MAD). A model should equip with long-term reasoning ability to ensure the image sequences are coherent in both event level and cinematic language level.

We present an encoder-decoder framework as a start point to overcome the above challenges for video storyboard generation. To improve the understanding of high-level text synopses, we finetune a pre-trained vision-language model (e.g., CLIP [25]) to retrieve relevant keyframes. Then an encoder-decoder framework with transformer architectures is proposed to auto-regressively predict image features, which can be used to retrieve and order images. However, it is difficult to train the decoder on a small dataset to retrieve coherent images. Inspired by the self-supervised pre-training paradigm in language modeling [4, 11], we consider the cinematic coherency is a visual language that is also possible to learn from large-scale unlabeled movie datasets. Therefore, we further propose a coherence-aware pre-training method that leverages large-scale movies without text synopses to pre-train the decoder.

We design two settings to evaluate methods: i) an ordering setting that provides models with oracle keyframes to re-order conditioning on the text, and ii) a retrieving-and-ordering setting that requires models to retrieve relevant frames from 500 candidate images and order them. Experimental results show that coherence-aware pre-training on unlabeled movies significantly improves the ordering performance . The larger the pretraining dataset, the better performance we could obtain in the downstream task. In addition, the time interval in pre-training videos matters. Both quantitative and qualitative results show that our model is able to create reasonable video storyboards. Nevertheless, there is still a long way to go to match the human perfor-

Table 1. Comparison between MovieNet-TeViS and other movie datasets.

Datasets	avgDuration	avg#Words	SecondsperWord	#unique bi-grams	avgConcreteness
LSMDC [32]	4.1s	9.03	0.4539	44.0K	2.993
MAD [36]	4.04s	12.69	0.3188	59.0K	2.991
CMD [1]	132s	18	7.33	83.2K	2.598
MovieNet-TeViS (Ours)	63.7s	24.82	2.82	134.5K	2.761

mance on this challenging TeViS task.

Our contributions are summarized as follows:

- We propose the **Te**xt Synopsis to **Vi**deo **S**toryboard task (**TeViS**) with the goal of retrieving an ordered sequence of images to visualize high-level text synopsis.
- We construct a **MovieNet-TeViS** benchmark based on MovieNet dataset [15]. It contains 10K text synopses with 4.6 keyframes on average for each synopsis.
- We establish an encoder-decoder baseline and propose Coherence-Aware Pre-training on Movies to improve coherence in long-term video storyboards.

2. Related Works

Our work is related to previous works of two categories: text-to-vision and movie understanding.

2.1. Text-to-Vision

Text-to-vision aims to retrieve or generate visual information corresponding to an input text. Inspired by the success of the pre-training paradigm in NLP [4, 11], recent advances in text-to-image retrieval also leverage massive image-text pairs to pre-train a large model for retrieval [10, 17, 19, 25, 45]. These methods achieve promising results on caption-based image retrieval tasks such as MSCOCO [9]. CLIP [22] adopts a dual-encoder architecture, uses 400 million image-text pairs for pre-training with a contrastive loss, and shows strong generalization power on cross-modal alignment. Some works also pre-train video-language models on large-scale video-text pairs [2, 23, 44]. However, text-to-image technologies can only produce static images which can not describe the dynamics in text synopsis, while text-to-video retrieval target searching existing video clips, rather than picking up keyframes from clips to collage out something new, which is demanded for if users input new texts. Text-to-vision generation also develops rapidly. Earlier methods widely adopt GAN-based methods conditioned on text [24, 31, 46]. Recent deep generative models use large Transformer networks [12, 27, 42] or diffusion models [26, 33] that can generate high-quality images. Text-to-video generation has been explored recently by extending advanced text-to-image generation methods [14, 35]. However, even advanced text-to-video generation methods can only generate GIF-like short videos without complicated motions and dynamics.

2.2. Movie Understanding

Existing works on movie understanding mainly explore content recognition and cinematic style analysis. Content recognition includes action [3, 21], scene recognition [6, 13, 29], and text-to-video retrieval task that focuses on movie datasets [1, 32]. Some works aim to analyze shot styles [28, 40], movie genre [34, 47] from a professional perspective. There are also some movie-related datasets [1, 15, 32]. LSMDC [32] dataset contains short clips paired with human-annotated captions. Condensed Movie Dataset (CMD) [1] consists of key scenes from the movie, each of which is accompanied by a high-level semantic description of the scene. The MAD [36] dataset is based on the LSMDC [32] dataset. Our constructed dataset is built on top of MovieNet [15] dataset, which is a large collection of movies annotated with many kinds of tasks such as scene segmentation, cinematic style classification, story understanding and so on. We use a subset of MovieNet annotated with text synopses of scenes. We manually construct video storyboard for the text synopses.

Tab. 1 presents the comparison of our dataset and related movie datasets. It shows that the duration of movie clips corresponding to a description in LSMDC and MAD is only 4 seconds and the average number of words in a description is only 9-12, which is much lower than ours. It is impossible to extract a meaningful storyboard from such short clips. Our MovieNet-TeViS and CMD give a synopsis or summary of 64-second or 132-second video segments respectively and thus we can expect such text is higher-level. Compared to CMD, our MovieNet-Tevis uses more words to describe video segments with half the duration of CMD. This indicates that our text synopses provide more details than those in CMD. As a start of such a challenging new task, our dataset is the most appropriate in duration of video clip and semantic level of text.

3. MovieNet-TeViS Dataset

Our goal is to assist amateur video makers to create video storyboards from text inputs. Since it is hard to obtain original video storyboards from professional video makers, we

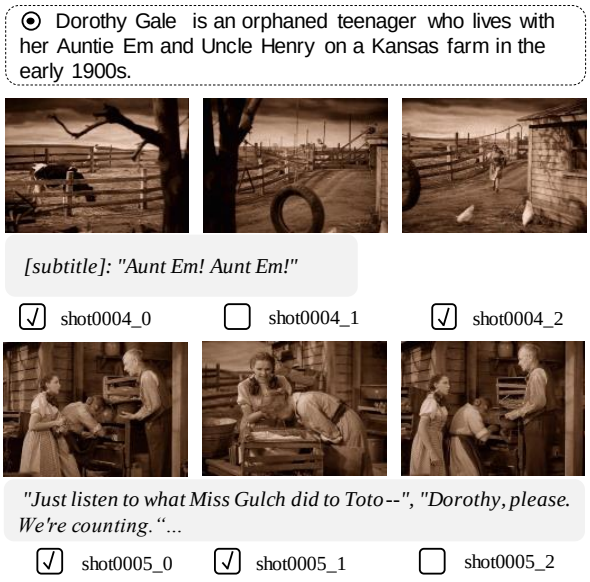


Figure 2. Annotating keyframes of a storyboard for a text synopsis

decided to select keyframes from released movies to reconstruct a succinct storyboard that a human user can use as a shooting plan. In terms of the text inputs, previous works have collected aligned script [49], caption [32], Descriptive Video Service (DVS) [38], book [48], or synopsis [37] to movies. However, books cannot be well-aligned with adapted movies; DVS is hard to obtain and thus limited in scale; wiki plots are too coarse, while scripts and captions are too detailed to compose for most non-professional users. We consider synopses are the most appropriate source which mimic the texts written by users in real scenario and contain desired level of details. MovieNet [15] and CMD [1] consist of such kind of synopses for movies. Although we could use the CMD dataset to scale up but we can expect that the CMD dataset would generate twice long sequence of images in a storyboard, which is much difficult to model, because it uses less words to describe the video clip with twice duration of our dataset. In the following, we first describe the dataset annotation process in Sec. 3.1 and then present analysis on our dataset in Sec. 3.2.

3.1. Data Annotation

MovieNet provides 4,208 text synopsis paragraph and movie segment pairs. A paragraph consists of 8 sentences (113 words) on average and a segmentation contains 95 shots with the variance of xx. It might be too difficult to learn semantic association and long-term reasoning over such long sequences with large variance. Therefore, we first split a paragraph into sentences and then align each sentence with a minimum number of keyframes in the movie segment.

Figure 3. Simplifying a storyboard by deleting redundant images

Fig. 2 shows the annotation interface. We present a synopsis paragraph sentence by sentence and all the shots aligned with the paragraph in MovieNet. For each shot, we display three evenly spaced frames as well as corresponding subtitles below the shot to help annotators understand the images better. The annotator should first select the sentences to form a text synopsis, and then choose a minimum number of images to visualize the text. We construct detailed guidelines to assure the quality of each annotated storyboard as follows:

1. The number of keyframes should be less than 20. Split the sentence if the number of selected keyframes is more than 20, or filter the sentence out of the dataset;
2. Do not select adjacent similar images, *e.g.*, for the example in Fig. 2, shot0004_1 should be deleted given shot0004_0;
3. The image must add value to express the synopsis sentence in terms of relevancy or coherency, *e.g.*, shot0005_2 cannot add any new value given shot0005_0;
4. If there is a basic conversation with a cycle of repeated images, only keep one pattern to make the storyboard succinct by selecting the first images in the first cycle and the last image in the last cycle. For example, in Fig. 3, A_iB_i is a basic conversation pattern and it has been repeated for 3 times. We ask annotators to select A_1 and B_3 to compose a complete conversation.

To ensure data quality, our data are annotated in three rounds. We hire 60 annotators in the first round to select keyframes that are necessary in relevancy or a vision language; in the second round the annotators further simplify or revise the storyboards by consistent rules; and six volunteer experts in the third round review and finalize the selected keyframes.

3.2. Dataset Analysis

Dataset statistics. Our collected MovieNet-TeViS dataset uses 2,949 paragraph-segment pairs from MovieNet after filtering improper examples by annotators. We sort storyboards by the number of keyframes ascendingly and use the

first 10,000 pairs of a synopsis sentence in English and a video storyboard, i.e., a sequence of keyframes as our final dataset. There are 45,584 keyframes in total. The number of keyframes in a storyboard ranges from 3 to 11 and about 60% storyboards consist of 3 or 4 keyframes. The average number of words in a synopsis sentence is about 24. In addition, MovieNet-TeViS covers 19 diverse movie genres. More details are presented in the supplementary material.

Concreteness measurement. The concreteness level of texts has a large influence on visualization difficulty. To systematically measure the concreteness of texts, we leverage a concreteness database introduced by Brysbaert et al [5] to calculate average concreteness of words in synopsis and compare with text descriptions of other datasets, i.e., LSMDC, MAD, and CMD and show the results in Tab. 1. To be specific, Brysbaert *et al*. [5] create a database that ask annotators to assign concreteness ratings from 1 to 5 for 40 thousand English words. The average ratings can evaluate the degree of how concrete a concept denoted by a word is. The larger value means more concrete. For example, the concreteness rating of "banana" is 5 while that of "love" is 2.07. As shown in the Tab. 1, our dataset has 2.76 average concreteness ratings while LSMDC and MAD have 2.99. This means that the text synopsis in MovieNet-TeViS is more abstract or higher-level than descriptions in LSMDC and MAD. CMD has 2.60 concreteness score which is slightly lower than ours. This makes sense because CMD use 18 words on average to describe 132-second video clips whereas our MovieNet-TeViS uses 24 words to describe 64-second video segments. As the first trial of a new task, our dataset has appropriate concreteness.

Diversity measurement. Following [41], we use the number of words, the number of unique n-grams and the number of words with different POS tags to compare diversity of text description or synopsis in LSMDC, MAD, CMD and our dataset. For fair comparison, we randomly sample 10,000 texts from LSMDC, MAD, and CMD datasets. We find that our built dataset MovieNet-TeViS has the richest n-grams, nouns, verbs, adjectives and adverbs. Due to space limitation, we only show the number of words and the number of unique bi-grams results in Tab. 1. We present the full comparison in the supplementary material. From the Tab. 1, we observe that CMD is also richer than LSMDC. This supports our observation that LSMDC has caption based description whereas CMD and MovieNet have high-level summary of movie clips or segments. Our dataset is richer than CMD, which is consistent with what the seconds per word shows. When looking into our dataset, we find that many text synopses contain dialogues, psychological descriptions, shot languages, etc. Such free text styles are closer to that of our target non-professional video makers.

4. Text Synopsis to Video Storyboard Task

The Text Synopsis to Video Storyboard (**TeViS**) task aims to retrieve a set of keyframes and order them to visualize the text synopsis. Assume we have a text synopsis $T = \{w_1, w_2, ..., w_n\}$ with n words, the goal of **TeViS** task is to retrieve m images from large candidate images and order them to visualize the text synopsis. The number of images m is different for each text synopsis T. We design two evaluation settings for the **TeViS** task: i) ordering the shuffled keyframes conditioned on the text, and ii) the task of retrieving then ordering.

4.1. Ordering the Shuffled Keyframes

Task. For a given text synopsis and its shuffled ground-truth images, how well can the models order them? This is a key step for creating a storyboard that needs to consider coherence across frames. To measure the long-term reasoning capability of models for ordering, we let the models order the ground-truth images for this evaluation.

Evaluation. For the ordering task, we are given a text synopsis and its shuffled ground-truth images, the models need to predict their order conditioned on text synopsis. We then can compute Kendall's τ [20] metric to report the result.

$$Kendall's\ \tau = 1 - \frac{2 * \#Inversions}{N * (N-1)/2} \tag{1}$$

where inversions are inverse-order pairs, i.e., the number of steps needed to switch to the original order. τ is always between -1 and 1, with 1 representing the full positive order and -1 representing the full inverse order.

4.2. Retrieve-and-Ordering Keyframes

Task. For a given text synopsis, how well can the models select the relevant images from a large set of candidates and then order them? This task is more practical in real situations.

Evaluation. For this evaluation, we are given a text synopsis and a large set of candidate images. The candidate images contain ground-truth images annotated by humans, and other negative images which are randomly sampled from other images in the corpus. The number of candidates including ground-truth and negative samples is 500. We consider both retrieval and ordering performance for this evaluation, thus we use the product of Recall@K and Kendall's τ as the final metric of this task. When some ground-truth images cannot be returned at top K, the Kendall's τ is calculated upon the returned ground-truth images at top K only.

5. Method

To provide a start point for tackling the task, we propose a text-to-image retrieval module based on a pre-trained

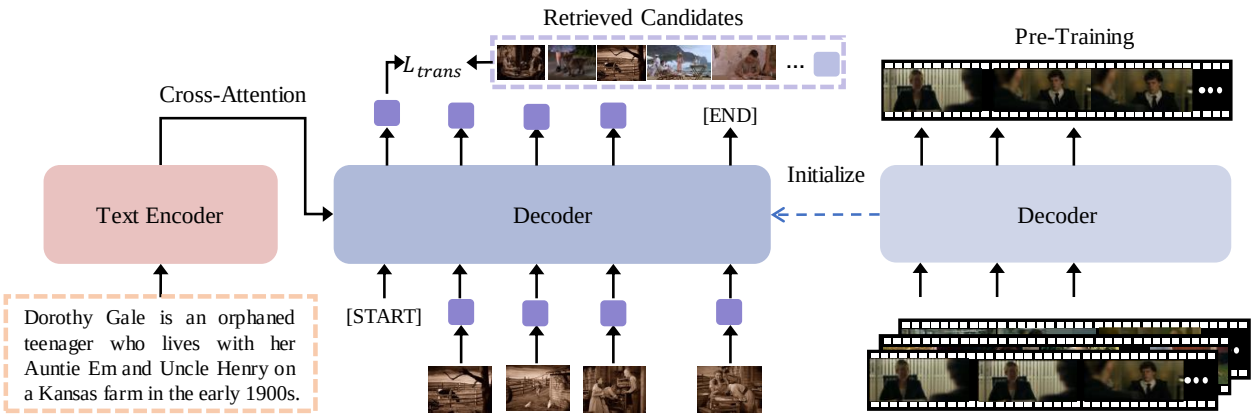


Figure 4. The framework of our Trans-TeViS Model for **TeViS** task.

image-text model (i.e., CLIP [25]), and an encoder-decoder module for ordering images. A coherence-aware pre-training method is further proposed to leverage large-scale movies to improve coherence across frames for the ordering module. We also present several strong baselines built on top of CLIP for ordering.

5.1. Text-to-Image Model for Retrieval

We leverage a pre-trained image-text model CLIP to conduct text-to-keyframe retrieval. During training, we randomly sample one frame from the ground-truth keyframe sequence to get positive image-text pair and frame from other sequences as negative for a text synopsis. Then we leverage a contrastive loss to maximize the similarity of matched images and texts while minimizing the similarity of unmatched images and texts, which is:

$$\begin{aligned}\mathcal{L}_{i2t} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(I_i^\top T_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(I_i^\top T_j/\tau\right)} \\ \mathcal{L}_{t2i} &= -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(T_i^\top I_i/\tau\right)}{\sum_{j=1}^{B}\exp\left(T_i^\top I_j/\tau\right)},\end{aligned} \tag{2}$$

where I_i and T_j are the normalized embeddings of i-th image and j-th sentence in a batch of size B and τ is the temperature. The overall text-image alignment loss $\mathcal{L}_{align}$ is the average of $\mathcal{L}_{i2t}$ and $\mathcal{L}_{t2i}$.

5.2. Encoder-Decoder Model for Ordering

Inspired by the encoder-decoder framework which is widely adopted in sequence generation tasks [9, 39, 43], we propose a **Trans-TeViS** model that adopts encoder-decoder architecture [39] to **Trans**late **Te**xt synopsis to **Vi**deo **S**toryboard. This design can not only sort the candidate images but also handle the variable length problem when creating video storyboards. As illustrated in Fig. 4, our model consists of a Transformer encoder E_T for text encoding, and a Transformer decoder D_T for image feature prediction. D_T predicts the image features auto-regressively with a cross-attention mechanism to condition on text. These predicted features can be used to retrieve images by dot-product similarity. The model is optimized with an NCE loss in each prediction step with negative images sampled randomly from a mini-batch:

$$\mathcal{L}_{trans} = -\frac{1}{BM}\sum_{i=1}^{B}\sum_{m=1}^{M}\log\frac{\exp\left(I_{i,m}^\top I_{i,m}/\tau\right)}{\sum_{I'\in\mathcal{N}_{i,m}\cup I_{i,m}}\exp\left(I_{i,m}^\top I'/\tau\right)} \tag{3}$$

where $I_{i,m}$ is the normalized embeddings of the m-th image from the i-th image sequence from the batch, $\mathcal{N}_{i,m}$ is the normalized embeddings of the negative images sampled from the batch.

Coherence-Aware Pre-training on Movies. Learning long-term reasoning for improving the coherence of re-ordered images is challenging, especially on a small dataset. It is hard for the model to learn sufficient movie-style evidence to make the transitions smooth across shots. Inspired by the success of the pre-training paradigm on NLP [4], which can learn language knowledge from massive data and then produce fluent sentences, we take a similar idea to leverage large-scale movies to learn the language of movies. Specifically, we pre-train the decoder part of our **Trans-TeViS** model with large-scale movie frame sequences without using text annotation. This method can be easily scaled up because movie frame sequences are easy to obtain.

5.3. Additional Baselines for Ordering

In addition to the proposed Trans-TeViS model, we design three strong baselines based on CLIP for ordering as shown in Fig. 5:

1) CLIP-Naive: we use CLIP to calculate the similarity between a text synopsis as query and its corresponding keyframes, and then order the keyframes based on the sim-

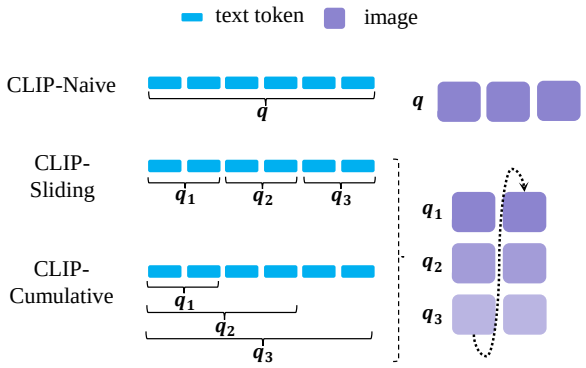


Figure 5. Illustration of additional baseline models for ordering.

ilarity scores.

2) CLIP-Sliding: we first divide the sentences into several segments as a group of queries where the number of segments is equal to the number of its corresponding keyframes. We then use sliding window to use each segment to retrieve the most similar keyframes in turn. Once a keyframe is chosen, this keyframe will be removed from the candidates.

3) CLIP-Cumulative: we first divide the sentences into several segments as CLIP-Sliding. However, when doing retrieval, we accumulate each segment and retrieve the most similar keyframes, which considers more context. For example, to retrieve the second keyframe, we use the first two segments as the query. We also remove the keyframes from the candidates once they are chosen in previous step.

6. Experiments

We evaluate the performance of proposed methods on MovieNet-TeViS dataset for the text synopsis to video storyboard task. We first describe the setup of experiment and then present the results of both ordering task and retrieve-and-ordering task. Finally, we show some qualitative results.

6.1. Experimental Setup

Implementation Details. We utilize CLIP-ViT-B/32 as the backbone in all compared methods. The initial learning rate is set to 1e-6, and we use a linear learning rate scheduler to decay the learning rate linearly after a warm-up stage. The network is optimized by AdamW optimizer, with the weight decay value of 5e-2 and the batch size of 16.

Pre-training datasets. We utilize the movies from CMD [1] dataset for pre-training the decoder of **Trans-TeViS** model. CMD dataset collects 7 to 11 clips with descriptions for each movie to cover the entire storyline. The CMD we exploited has 30K clips in the training set, 2K clips in the validation set, and 1K clips in the test. To balance between information richness and computational complexity, we use a uniform frame sampling strategy to extract 5 frames every clip for pre-training .

Table 2. Results of ordering task. Our method Trans-TeViS achieves the best performance, though still leaving much room for improvement compared to human capabilities. $[s-e]$ under Kendall's τ denotes sequence length from s to e.

Method	Kendall's $\tau\uparrow$		
	all	[3-5]	[6-11]
Human	0.821	0.860	0.734
CLIP-Naive	0.183	0.248	0.036
CLIP-Sliding	0.230	0.278	0.123
CLIP-Cumulative	0.244	0.291	**0.139**
Trans-TeViS	**0.261**	**0.324**	0.120

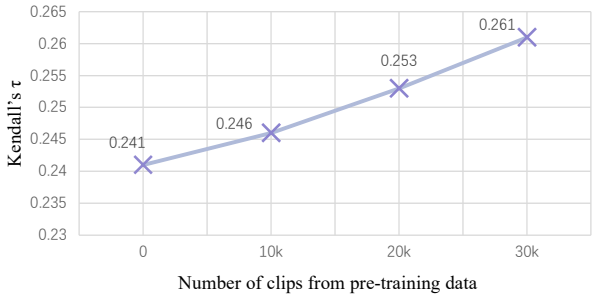


Figure 6. Ablation study results of the pre-training dataset with different scales. The performance of our method improves as the scale of the pre-training data increases.

6.2. Ordering Task

Results. We conducted several experiments to verify the effect of different methods on the text synopsis to video storyboard task, and results are shown in Tab. 2. The CLIP-Naive method achieves the poorest performance due to the lack of sequence modeling. The CLIP-Sliding and CLIP-Cumulative methods outperform the CLIP-Naive method, proving to be more effective ways of applying CLIP, thanks to the ability to segment semantic information of the text and thus able to model sequences. The Trans-TeViS method achieves the best overall result of all the models, demonstrating the ability of sequence generation models to learn long-term information in keyframe sequences.

In addition, a human study was conducted to make a better assessment of the task. We invited participants to reorder the shuffled keyframe sequences. The performance of humans is presented in Tab. 2. Humans achieve much better performance than our best model, suggesting there is high potential for improvement.

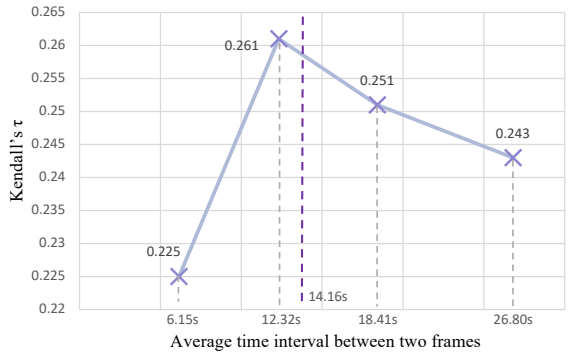


Figure 7. Ablation study results of pre-training dataset with different average time interval between every two frames. Pre-training dataset with average time intervals similar to the MovieNet-TeViS (purple line of 14.16s) leads to a better performance.

Table 3. Text-to-image retrieval performance on MovieNet-TeViS. We compare CLIP models with and without fine-tuning (ft).

Method	R@1↑	R@5↑	R@10↑	R@50↑
CLIP w/o ft	5.73	19.72	28.98	56.42
CLIP	7.48	26.34	38.94	68.90

Analysis of Pre-training Data. To verify the impact of dataset scales in pre-training, we randomly sample subsets with different sizes from the CMD dataset for pre-training which has 30K clips originally. As shown in Fig.6, the method without any pre-training obtains the poorest performance, indicating the importance of Coherence-Aware Pre-training. It can be seen that the performance of the method improves as the scale of the pre-training dataset increases.

Analysis of time interval. In the original CMD dataset, we use a uniform sampling strategy to extract 5 frames for each clip, and the average time interval (i.e. 1/fps) between every two frames is 26.8s, while in MovieNet-TeViS this number is 14.16s. To explore the impact of this gap, we extract frame sequences with the same sequence length and different time intervals from CMD as pre-training data. As shown in Fig.7, pre-training data with average time intervals similar to the MovieNet-TeViS leads to a better performance.

6.3. Retrieve-and-Ordering Task

Retrieval Results. We first evaluate the performance of text-to-image retrieval. We compare CLIP models [25] with and without fine-tuning on our MovieNet-TeViS dataset. As shown in Tab. 3, even without fine-tuning on our dataset, CLIP shows reasonable performance. The fine-tuned CLIP model can achieve much better performance.

Table 4. Results of Retrieve-and-Order Task. Our method Trans-TeViS outperforms other CLIP-based methods.

Method	Kendall's τ↑				
	R@10	R@20	R@30	R@40	R@50
CLIP-Naive	0.223	0.169	0.133	0.143	0.152
CLIP-Sliding	0.218	0.204	0.189	0.208	0.246
CLIP-Cumulative	0.226	0.203	0.188	0.214	**0.249**
Trans-TeViS	**0.276**	**0.271**	**0.253**	**0.251**	0.244

Retrieve-and-Ordering Results. We report the result of the Retrieve-and-Ordering Task in Tab. 4. It can be seen that the method of Trans-TeViS achieves the best performance while CLIP-Naive has the poorest, with CLIP-Sliding and CLIP-Cumulative in the middle. This result is consistent with the experimental result of the ordering task, suggesting that the difference in performance primarily stems from the difference in methods' ability of ordering.

6.4. Qualitative Results

In addition to the quantitative results, we further carry out a case study on how well our proposed methods perform in the TeViS task. As Fig. 8 shows, for the given text synopsis, our proposed Trans-TeViS method performs the best and can correctly order No.1, 3, and 4 keyframes and No.2, 3, and 4 ones. CLIP-Naive takes the synopsis as the whole to encode and thus it actually considers relevance only without any order information. It performs worst as expected. Our proposed CLIP-Sliding and CLIP-Cumulative address the limitation because text synopsis is split into several text fragments and the ordering of keyframes depends on the ordering of text fragments. In this case, the text fragments are well aligned with ground-truth keyframes from human's perspective, but it is still difficult for CLIP-Sliding and CLIP-Cumulative in ordering No.2 and 3. Our proposed pre-training and transformer based model can correctly order No.2 and 3, which shows the advantages in learning the visual language for storyboard creation.

7. Conclusion

In this paper, we introduce a novel **TeViS** task (**Te**xt synopsis to **Vi**deo **S**toryboard), which aims to retrieve an ordered sequence of images to visualize the text synopsis. We also construct a **MovieNet-TeViS** dataset to support the task. To align the diverse text synopsis with keyframes, we utilize a pre-trained Image-Text model to overcome this challenge. We propose an encoder-decoder model called **Trans-TeViS** which translates text synopsis to keyframe sequence. We also propose Coherence-Aware Pre-training on Movies to improve the long-term reasoning of the de-

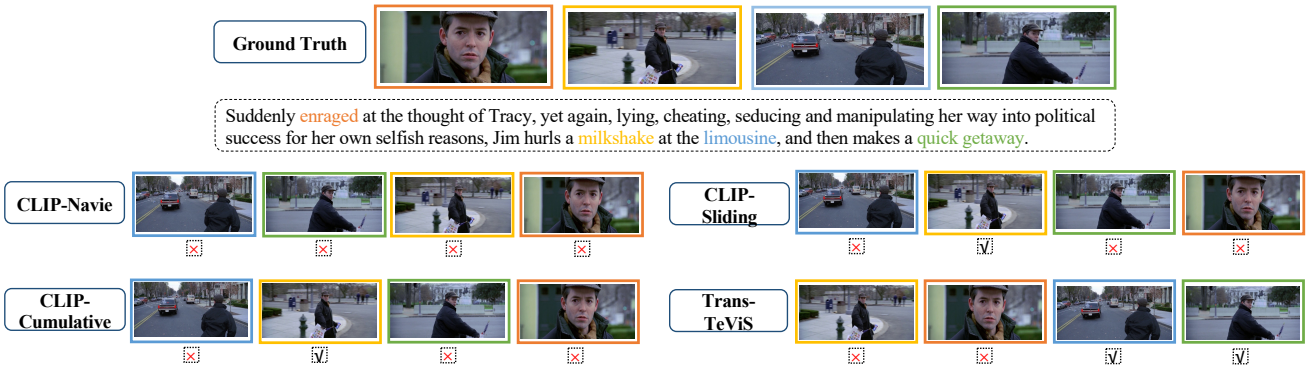


Figure 8. Qualitative examples of different models for the ordering task on our Movie-TeViS dataset.

coder for ordering the keyframes. Ablation studies verify the effectiveness of our proposed model. Both quantitative and qualitative results show our method is better than other baselines.

References

[1] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Computer Vision – ACCV 2020: 15th Asian Conference on Computer Vision, Kyoto, Japan, November 30 – December 4, 2020, Revised Selected Papers, Part V*, page 460–479, Berlin, Heidelberg, 2020. Springer-Verlag. 1, 3, 4, 7

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 1, 3

[3] Piotr Bojanowski, Francis R. Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Finding actors and actions in movies. *2013 IEEE International Conference on Computer Vision*, pages 2280–2287, 2013. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 2, 3, 6

[5] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46:904–911, 2014. 5

[6] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. Scene detection in videos using shot clustering and sequence alignment. *IEEE transactions on multimedia*, 11(1):89–100, 2008. 3

[7] Shizhe Chen, Bei Liu, Jianlong Fu, Ruihua Song, Qin Jin, Pingping Lin, Xiaoyu Qi, Chunting Wang, and Jin Zhou. Neural storyboard artist: Visualizing stories with coherent image sequences. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 2236–2244, New York, NY, USA, 2019. Association for Computing Machinery. 1

[8] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10638–10647, 2020. 1

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 3, 6

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer, 2020. 1, 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*, pages 4171–4186, 2019. 2, 3

[12] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 3

[13] Bo Han and Weiguo Wu. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In *2011 IEEE International conference on multimedia and expo*, pages 1–6. IEEE, 2011. 3

[14] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1, 3

[15] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 4

[16] Ting-Hao K. Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Aishwarya Agrawal, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, 2016. 1, 2

[17] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, pages 12976–12985, 2021. 3

[18] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 1

[19] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. Wenlan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021. 1, 3

[20] Mirella Lapata. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 2006. 5

[21] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 3

[22] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, pages 7331–7341, 2021. 1, 3

[23] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, pages 2630–2640, 2019. 1, 3

[24] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1789–1798, 2017. 3

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 6, 8

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[27] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 3

[28] Anyi Rao, Jiaze Wang, Linning Xu, Xuekun Jiang, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A unified framework for shot type classification based on subject centric lens. In *European Conference on Computer Vision*, pages 17–34. Springer, 2020. 3

[29] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10143–10152, 2020. 3

[30] Hareesh Ravi, Lezi Wang, Carlos Manuel Muñiz, Leonid Sigal, Dimitris N. Metaxas, and Mubbasir Kapadia. Show me a story: Towards coherent neural story illustration. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7613–7621, 2018. 1

[31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 3

[32] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 1, 2, 3, 4

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 3

[34] Gabriel S Simões, Jônatas Wehrmann, Rodrigo C Barros, and Duncan D Ruiz. Movie genre classification with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 259–266. IEEE, 2016. 3

[35] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 3

[36] Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. 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, 2022. 2, 3

[37] Yidan Sun, Qin Chao, and Boyang Li. Synopses of movie narratives: a video-language dataset for story understanding. *arXiv preprint arXiv:2203.05711*, 2022. 4

[38] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070*, 2015. 4

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 6

[40] Hee Lin Wang and Loong Fah Cheong. Taxonomy of directing semantics for film shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 19:1529–1542, 2009. 3

[41] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 5

[42] Chenfei Wu, Jian Liang, Xiaowei Hu, Zhe Gan, Jianfeng Wang, Lijuan Wang, Zicheng Liu, Yuejian Fang, and Nan

Duan. Nuwa-infinity: Autoregressive over autoregressive generation for infinite visual synthesis. *arXiv preprint arXiv:2207.09814*, 2022. 1, 3

[43] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 6

[44] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022. 1, 3

[45] Hongwei Xue, Yupan Huang, Bei Liu, Houwen Peng, Jianlong Fu, Houqiang Li, and Jiebo Luo. Probing intermodality: Visual parsing with self-attention for vision-and-language pre-training. In *NeurIPS*, volume 34, pages 4514–4528, 2021. 1, 3

[46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017. 3

[47] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 747–750, 2010. 3

[48] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015. 4

[49] Yutao Zhu, Ruihua Song, Jian-Yun Nie, Pan Du, Zhicheng Dou, and Jin Zhou. Leveraging narrative to generate movie script. *ACM Transactions on Information Systems (TOIS)*, 40(4):1–32, 2022. 4