Synchronized Video Storytelling: Generating Video Narrations with Structured Storyline

Dingyi Yang[1], **Chunru Zhan**[2], **Ziheng Wang**[1], **Biao Wang**[2], **Tiezheng Ge**[2], **Bo Zheng**[2], **Qin Jin**[1]*
[1]Renmin University of China
[2]Alibaba Group
{yangdingyi,zihengwang,qjin}@ruc.edu.cn
{zhanchunru.zcr,eric.wb,tiezheng.gtz,bozheng}@alibaba-inc.com

Abstract

Video storytelling is engaging multimedia content that utilizes video and its accompanying narration to attract the audience, where a key challenge is creating narrations for recorded visual scenes. Previous studies on dense video captioning and video story generation have made some progress. However, in practical applications, we typically require synchronized narrations for ongoing visual scenes. In this work, we introduce a new task of Synchronized Video Storytelling, which aims to generate synchronous and informative narrations for videos. These narrations, associated with each video clip, should relate to the visual content, integrate relevant knowledge, and have an appropriate word count corresponding to the clip's duration. Specifically, a structured storyline is beneficial to guide the generation process, ensuring coherence and integrity. To support the exploration of this task, we introduce a new benchmark dataset E-SyncVidStory with rich annotations. Since existing Multimodal LLMs are not effective in addressing this task in one-shot or few-shot settings, we propose a framework named VideoNarrator that can generate a storyline for input videos and simultaneously generate narrations with the guidance of the generated or predefined storyline. We further introduce a set of evaluation metrics to thoroughly assess the generation. Both automatic and human evaluations validate the effectiveness of our approach. Our dataset, codes, and evaluations will be released.

1 Introduction

Video storytelling (Li et al., 2019; Bhattacharya et al., 2023) is a tactic that utilizes the naturally engaging video format to share stories and has been widely used as a powerful tool for marketing, educational, or entertainment purposes. The video and its accompanying narration are both critical to successful video storytelling. Therefore, two aspects of automatic video storytelling have been explored: generating or retrieving videos to illustrate written stories (He et al., 2023; Lu et al., 2023); and generating narrations based on visual scenes (Li et al., 2019). Nowadays, people frequently use convenient devices to make visual recordings. These initial recordings can be transformed into compelling video stories by integrating synchronized narrations. There is a strong need for automatically generating narrations for a given video.

Figure 1: An example of our annotated synchronized video storytelling, with English translations provided for easy reading. Sequential visual scenes are narrated, following a storyline expressed by categorized script labels. Each narration should incorporate appropriate relevant knowledge (as underlined in the figure) and have a word count that fits the duration of each clip.

Existing research on video-to-text generation mostly focuses on creating a single sentence to summarize the video (Wang et al., 2019) or generating fine-grained captions for video clips (Krishna et al., 2017). Gella et al. (2018) and Li et al. (2019) take a step further by generating sequential visual sum-

*Corresponding Author.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9479–9493

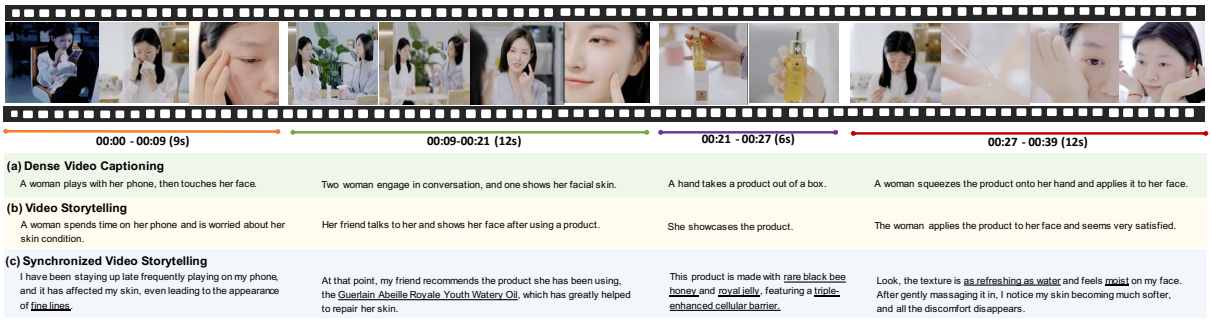

Figure 2: Comparison of the proposed Synchronized Video Storytelling and existing video-to-text generation tasks. (a) Dense Video Captioning aims to generate corresponding narrations for multiple events within a video. (b) Video Storytelling aims to form a coherent story with these narratives. (c) The proposed Synchronized Video Storytelling aims to generate synchronized, informative, and coherent narrations to support voiceovers for videos.

maries for video scenes, creating a coherent video story. However, in real-life applications, when narrating an ongoing video, each narration should not only reflect the current visual scene, but also have an appropriate word sequence length to fit the duration of the video clip. Including relevant external knowledge is also important, as it can attract and convince the audience, particularly for marketing purposes. Additionally, we believe that a logical storyline helps maintain the coherence and integrity of the narratives.

In this paper, we introduce the task of **Synchronized Video Storytelling**, which aims to generate synchronized narrations for videos. As illustrated in Figure 1, the sequential visual scenes in the video are accompanied by corresponding narrations. Each narration has a suitable word sequence length and incorporates relevant knowledge to effectively engage the audience. They can be directly utilized as voiceovers for video recordings. This distinguishes our task from existing ones, as illustrated in Figure 2. Furthermore, our generation follows a structured storyline, similar to the human writing process, beginning with a brief outline and then drafting the complete story (Yang et al., 2022b). This proposed task is complex and cannot be effectively solved by traditional frameworks or evaluated by existing benchmarks. A possible approach is to utilize powerful Multi-modal Large Language Models (MLLMs) (Han et al., 2023; Sun et al., 2023) for zero-shot or few-shot generation. However, most MLLMs face challenges with time alignment and sequential video comprehension (Huang et al., 2023). Moreover, generating narrations for longer videos becomes difficult for MLLMs due to the requirement of excessively long visual tokens. Another approach is to generate captions for video shots and use them as input for the LLMs (Bhattacharya et al., 2023; Lin et al., 2023b). However, this relies heavily on the performance of the captioning model. Furthermore, current models are ineffective in generating coherent, appealing, and length-appropriate narrations. Additional training on specific instruction datasets is necessary.

To support the exploration of the proposed task, we create a benchmark dataset named E-SyncVidStory (**E-**commerce **Sync**hronized **Vid**eo **Story**telling). Given the rapid growth of e-commerce platforms, many sellers have product information and visual materials, but may lack the time and resources to convert these into engaging advertisement videos. Therefore, they seek an automatic and efficient solution. Our dataset can provide value for both research and practical applications in this area. E-SyncVidStory contains 6.0K videos with 41.3K video clips in total. Each video is segmented as sequential visual scenes with synchronized Chinese narrations. As the primary objective of Ad stories is to engage with the audience and showcase products (Bhattacharya et al., 2023), they might not always present a sequence of events as a classic story. Still, like classic ones, advertising stories should be coherent and follow a logical storyline, which is annotated in our dataset[1]. All annotations are obtained through automatic preprocessing, GPT-based refinement, and manual checks.

We further propose an effective framework integrating vision models and LLMs to address the task of synchronized video storytelling. Our specifically designed instruction structure allows simultaneous

[1]The storyline is expressed by categorized script labels. These labels are divided into 12 types, with their definitions provided in Appendix A.4.

generation of the storyline and the synchronized story. To enhance the effectiveness of visual embedding, we compress the long visual tokens into shorter visual information, which includes the long-term memory of previous clips and compressed information of the current clip. To evaluate the generated stories, we propose a set of evaluation metrics to comprehensively measure the generated results. Extensive experiments on both specific domain (E-SyncVidStory) and general domain (Li et al., 2019) Video Storytelling verify the effectiveness of our proposed framework.

The main contributions of our work include: 1) We introduce the new task of synchronized video storytelling, which is more challenging as it requires generating a synchronous, informative, and coherent story for a given video. 2) To support the exploration of this task, we collect a benchmark dataset in the advertising domain, namely E-SyncVidStory, which contains rich annotations, and also can support more research beyond video storytelling; 3) We propose an effective framework called VideoNarrator, which takes relevant knowledge and sequential video scenes as inputs. It simultaneously supports storyline generation and controllable video story generation; 4) We introduce a set of systematic evaluation metrics to comprehensively measure the story generation results. Automatic and human evaluation results verify the effectiveness of our method.

2 Related Works

Video Narration Generation. Bridging vision and natural language is a longstanding goal in computer vision and multimedia research (Li et al., 2019). Previous research on video-to-text generation primarily focuses on video captioning tasks, generating single-sentence factual descriptions for the whole video (Wang et al., 2019; Xu et al., 2016). Some works aim to provide more detailed and comprehensive descriptions by generating multi-sentence paragraphs (Krishna et al., 2017; Zhou et al., 2018a). However, most of these works ignore the overall coherence of the video descriptions. Gella et al. (2018) and Li et al. (2019) step further to generate a coherent and succinct story from abundant and complex video data. However, these datasets only generate short summaries for long videos and do not handle synchronized video storytelling. Additionally, all these works only focus on the visual content and do not take into account the need to incorporate external knowledge.

Multi-modal Large Language Models. With the success of Large Language Models (LLMs), many works have tried to build Multi-modal LLMs by combining models from other modalities. In the video field, VideoChat (Li et al., 2023b) integrates video foundation models and LLMs via a learnable neural interface, excelling in spatiotemporal reasoning, event localization, and causal relationship inference. VideoChat Longvideo (OpenGVLab) incorporates LangChain into VideoChat to support videos that are longer than one minute. Video-ChatGPT (Maaz et al., 2023a) design a 100k video instruction dataset, successfully empowering LLMs to comprehend videos. Video-LLaVA (Lin et al., 2023a) learns from a mixed dataset of images and videos, and performs joint visual reasoning on images and videos. However, all of these methods are not effective in comprehending sequential video clips. VtimeLLM (Huang et al., 2023) design a boundary-aware three-stage training strategy, enhancing results for dense video comprehension. However, it struggles with longer video inputs since the maximum training length is 2048. Like other MLLMs, it also lacks efficiency in generating video descriptions that are constrained by appropriate text length and specific storylines.

3 Synchronized Video Storytelling

3.1 Task Definition

Given a video V composed of sequential video clips $\{v_1, v_2, ..., v_n\}$ and the related knowledge composed of several knowledge items $\mathcal{K} = \{k_1, k_2, ..., k_m\}$, synchronized video storytelling aims to incorporate the knowledge to generate a narration s_i for each video clip v_i, and all narrations form a coherent story $\mathbf{S} = \{s_1, s_2, ..., s_n\}$ for the video. The word count [2] of each s_i should be within the appropriate range r_i. To enhance the coherence of the overall story, the generation is guided by a predefined or automatically predicted storyline, which is expressed using a sequence of script labels $\{l_1, l_2, ..., l_n\}$. Each script label l_i, belonging to one of the 12 predefined types, summarizes the key point of the corresponding narration. More details about the script label are presented in the Appendix A.4.

[2] If the time length of video clip v_i is t seconds, considering the normal speaking speed of 5 words per second, the appropriate word length range can be $[5t - 3, 5t + 3]$ with an error margin of 3 words in either direction.

Table 1: Comparison between E-SyncVidStory and existing dense video captioning (top half) and video story generation (lower half) datasets.

	Input Modality	Storyline	Num. videos	Avg. clips (per video)	Avg. text len. (per video)	Avg. text len. (per second)
YouCook2 (Zhou et al., 2018b)	Video	-	2K	7.7	67.7 (en)	0.21
ActivityNet Captions (Krishna et al., 2017)	Video	-	20K	3.7	47.6 (en)	0.40
Charades-STA (Gao et al., 2017)	Video	-	10K	1.8	11.0 (en)	0.36
ViTT (Huang et al., 2020)	Video	-	8K	5	110.5 (en)	0.44
VideoStory (unreleased) (Gella et al., 2018)	Video	-	20K	6.1	-	-
Video Storytelling (Li et al., 2019)	Video	-	105	13.5	162.6 (en)	0.22
E-SyncVidStory	Video & Knowledge	yes	6K	6.9	194.1 (zh)	5.21

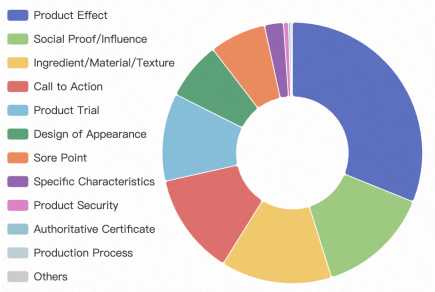

Figure 3: The distribution of the number of 12 types of script labels in the benchmark.

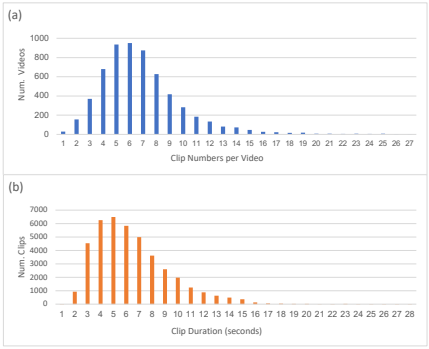

Figure 4: Distribution of clip numbers per video and the clip duration.

3.2 E-SyncVidStory Dataset

Data Collection. We select high-quality advertisement videos with high click-through rates from the web and collect these videos along with their related information which will be utilized as knowledge points. As shown in Figure 1, knowledge points are some short expressions about different attributes of the product. We automatically divide each video into sequential video clips (Yue Zhao, 2019) and obtain their synchronized narrations through the automatic speech recognition (ASR) tool (Zhang et al., 2022). However, these pre-processed narrations still contain errors. To minimize the labor costs to correct them, we utilize the powerful tool GPT-4 (OpenAI, 2023) to correct the errors and predict script labels for each narration. The prompts for ASR refinement and script label classification are displayed in our Appendix A.6. With the aforementioned automatic pre-processing process, the results are already quite satisfactory. We then recruit crowd workers to review the narrations, check script label classification, and correct any remaining errors. This way, we build our proposed dataset, named **E-SyncVidStory**.

Statistics and analysis. E-SyncVidStory contains 6,032 videos, with a total of 41,292 video clips. The advertising videos cover various industries including personal care, makeup, clothing, household supplies, maternal parenting, and electronics, etc. The video length ranges from 2 to 220 seconds with an average of 39 seconds, and the average story length is 194 words. As shown in Figure 4 (a), the clip numbers within each video vary greatly. Thus we apply relative clip position embedding as described in Section 4.1. The duration of video clips is displayed in Figure 4 (b). Since a single clip within a video can have a maximum duration of 30 seconds, visual compression is needed when comprehending the video. The distributions of script labels are shown in Figure 3. Most narrations focus on the product effect, demonstrating the product's benefits. They also prefer social proof, using positive experiences to persuade the audience. While this work focuses on creating a story from a structured video, this dataset could also aid investigation in generating a well-structured video from multiple unsequential clips, as discussed in Appendix A.5.

Dataset Comparison. To the best of our knowledge, E-SyncVidStory is the first synchronized video storytelling benchmark. As shown in Table 1, the average text length per second is much longer than the previous datasets because it allows for synchronous narration. We also annotate the structured storyline and corresponding knowledge of videos, which can be used to generate more informative and coherent narrations.

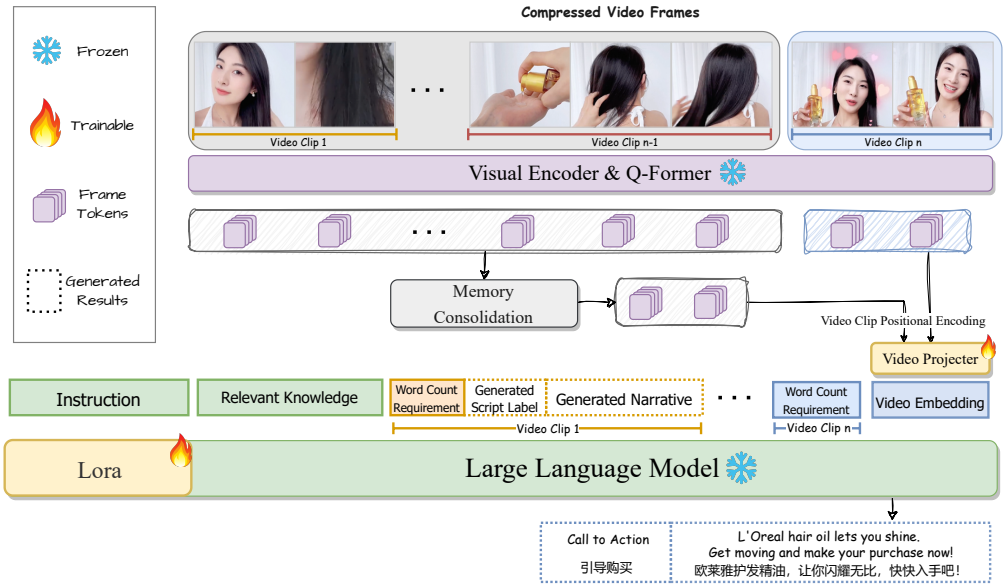

Figure 5: Illustration of our proposed VideoNarrator. The architecture is based on the Visual Feature Extractor, the Video Projector, and the LLM. The original video is compressed, retaining only the important frames. When generating the narration for video clip n, the previous frames are combined into fixed-length video memory tokens and concatenated with the current frame tokens. We pass the concatenated visual features and the relative video clip position embedding through the video projector to achieve the final video embedding. Based on the visual embedding and the previously generated narrations, the LLM will generate an appropriate script label and apply it to guide the narration generation.

3.3 Automatic Evaluation Metrics

Following Li et al. (2019), we evaluate the generated narrations with NLP metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and CIDEr (Vedantam et al., 2015). However, it is not reasonable to only consider the scores based on Ground Truth, as the generated story may include different knowledge even from the same aspect. For example, while the visual scene displays the list of ingredients, the narration might select a portion of them to describe. Thus we propose reference-free metrics to evaluate the results, in the aspect of visual relevance, knowledge relevance, controllable accuracy, and fluency.

Visual Relevance. Following Shi et al. (2022), we use EMScore and EMScore$_{\text{ref}}$ to evaluate the visual relevance between the video and generated narrations. The textual and visual embeddings for evaluation are encoded using Chinese-Clip (Yang et al., 2022a). Unlike the traditional visual relevance evaluation, we only consider the similarity between visually relevant words[3] and the video clips. Because connecting words and knowledge within a sentence which enriches the video narrations should not hurt the visual relevance even if they have a low similarity score.

Knowledge Relevance. A compelling advertising story should incorporate relevant knowledge to effectively attract customers. Firstly, each narration should incorporate relevant knowledge points while avoiding unrelated messages, denoted as information similarity (Info$_{\text{Sim}}$). Additionally, the knowledge points within a story should be diverse, avoiding the emphasis on only a small amount of information. This is referred to as information diversity (Info$_{\text{Diverse}}$). The detailed evaluation process is displayed in our Appendix A.2.

Controllable Accuracy. We evaluate the accuracy of the word sequence length by checking if the word count of sub-story s_i falls within the appropriate range as described in Section 3.1. When generating stories controlled by a predefined storyline, we utilize GPT-4 to assess the controllable accuracy of the script labels. The prompt for this evaluation can be found in Appendix A.6.

Fluency. We apply intra-story repetition (Yao et al., 2019) which measures sentence repetition within a story through word overlaps.

[3]We tokenize s_i using jieba (Sun, 2019) and retain only the nouns, verbs, and adjectives as visually relevant words.

4 Method

As shown in Figure 5, our proposed VideoNarrator consists of the visual feature extractor, the memory consolidation component, the video projection layer, and the Large Language Model (LLM). Inspired by Maaz et al. (2023b), we utilize the powerful CLIP-L/14 visual encoder (Radford et al., 2021) and the Q-Former module from BLIP-2 (Li et al., 2023a) to obtain frame-level visual features. These features already achieve a good alignment with the textual modality. To capture temporal information, we further feed them into a linear video projector. However, inputting all video frames would result in significant computational complexity and memory usage (Song et al., 2023). Moreover, it would make it challenging to extract prior knowledge or maintain coherence with the previously generated story. To address these difficulties, we consolidate the visual information from previous clips into fixed-length memory tokens. Additionally, we compress the visual frames of the current clip, retaining only the key information.

4.1 Visual Embedding

Visual Compression. Considering the visual information redundancy (Tong et al., 2022), we first compress the original frames in each video clip by removing a frame from any adjacent pair frames if their similarity is above a threshold value τ. This removal process will be continued until the similarity between each pair of adjacent frames falls below τ.

Memory Consolidation. For long videos consisting of multiple video clips, when generating a narration for the current clip v_i, we only need a more concise message of previous video clips. Inspired by Song et al. (2023), we perform memory consolidation by merging the most similar visual tokens. Specifically, we maintain fixed-length tokens as visual memory. We iteratively repeat the merging process [4] until the memory tokens of previous video clips reach the fixed length.

Video Clip Positional Encoding. The relative position of each video clip is important in video story generation. For example, when approaching the end of a video, it is more likely to evoke a response from the audience rather than introduce the product. However, the model could only acknowledge the absolute position of a video clip. Thus, we propose a relative clip position embedding and add it to the video embeddings. If a clip is located at $p\%$ of the whole video, then p is its relative position. This value p is passed through a positional embedding layer, whose parameters are updated during training. The encoded position feature is then added to the visual embeddings.

[4] Assume $\{f_i\}_{i=1}^{k}$ are the visual features of k frames. In each merging process, we : (1) Evaluate the cosine similarity of each adjacent pair. (2) Find the pair (f_m, f_{m+1}) with the highest similarity. (3) Merge f_m and f_{m+1} by averaging their features and remove f_{m+1}.

System Role:
你是一位广告商，擅长为商品撰写具有吸引力的广告文案，并且能够准确理解商品信息和视频信息。

Instruction:
我将为你提供商品信息以及该商品对应的广告视频的每个镜头的信息。你的任务是结合商品信息，为视频的每个镜头撰写符合当前字数要求的广告文案，吸引消费者购买。你需要判断每个镜头的剧本类型，确保文案围绕镜头的剧本类型展开，与当前场景紧密相关，并且与前面的文案保持连贯。
请注意，在撰写文案时，主要使用到提供的商品信息以及镜头的场景信息。

User:
商品信息:
{所有相关信息}

视频:
镜头1:
字数要求: 22-28个字
剧本类型: 痛点铺垫
文案: 1.发质真的很影响一个人的外形，可惜我的头发天生毛躁。
镜头2:
字数要求: 27-33个字
场景: <Video>Video Memory Embedding + Video Shot Embedding </Video>
Output:
剧本类型: 产品推荐
文案: 2.于是我开始寻找好的精油，最后入手了这款欧莱雅护发精油。

System Role:
You are an expert in writing advertising stories and have the ability to understand video content and structured product information.

Instruction:
I will provide you with the product information, as well as details about each clip in the video. Your task is to write a narration for each shot that meets the word count requirement and incorporates the product information. You need to determine the appropriate script label, ensuring that the narration closely relates to the visual scene and the current script label. Each narration should remain coherent with the previously generated text. Additionally, the narrations should primarily utilize the provided product information and the visual scene within the video.

User:
Product Information:
{The knowledge points}
Video:
Video Clip 1:
Word count requirement: 22-28 words
Script label: Sore Point
Narration: 1. Hair texture can greatly affect a person's appearance, but unfortunately, my hair is naturally frizzy.
Video Clip 2:
Word count requirement: 27-33 words
Visual scene: <Video>Video Memory Embedding + Video Shot Embedding </Video>
Output:
Script label: Social Proof
Narration : 2. So I began searching and ended up with L'Oreal Hair oils.

Figure 6: This figure illustrates our system prompt, with English translations provided for easy reading. The instructions are designed to help the model understand the task of synchronized video storytelling. Here, we present the constructed training sample for the second clip within a video.

4.2 Prompting

The input prompt for VideoNarrator is displayed in Figure 6. For each video clip, the controllable signal is provided just before the narration, which can ensure controllable accuracy. In our instruction, we concisely explain the task requirements. As shown in Figure 6, these red words can make it

Table 2: Automatic Evaluation Results on E-SyncVidStory. The best result on each aspect is **bolded**.

		CIDEr	METEOR	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Multi-Model Pipelines	LLaVA-1.5+GPT-3.5 (Zero-Shot)	15.0	8.5	18.4	7.6	4.3	2.7
	LLaVA-1.5+GPT-3.5 (Few-Shot)	18.6	8.3	20.0	8.5	4.9	3.2
End2end MLLMs	Video-ChatGPT (Maaz et al., 2023a)	4.1	8.6	13.8	5.5	2.9	1.7
	Video-LLaVA (Lin et al., 2023a)	6.1	8.3	15.1	6.0	3.3	1.8
	VTimeLLM (Huang et al., 2023)	8.4	8.6	14.9	6.5	3.9	2.6
Fine-tuned End2end MLLMs	VTimeLLM (w/ finetune)	28.0	8.6	19.8	9.3	5.6	3.8
	VideoNarrator (Ours)	**33.3**	**9.1**	**21.0**	**9.9**	**6.3**	**4.4**
	a. w/o LLM LoRa finetune	10.5	6.9	16.5	5.4	2.8	1.7
	b. w/o Storyline	26.2	8.2	19.4	9.4	5.9	3.6
	c. w/o Video Clip Position	29.5	8.7	20.1	9.1	5.6	3.9
	d. w/o Visual Compression	29.6	8.7	20.4	9.2	5.7	3.8
	e. w/o Visual Memory	22.3	8.2	19.2	8.1	4.7	3.1

Table 3: Automatic Evaluation Results on E-SyncVidStory, considering four aspects (Section 3.3): visual relevance, knowledge relevance, controllable accuracy, and fluency. The best result on each score is **bolded** and the second best is underlined.

	Visual Relevance		Knowledge Relevance		Controllable Acc.	Fluency
	EMScore↑	EMScore $_{\text{ref}}$ ↑	Info$_{\text{Sim}}$ ↑	Info$_{\text{Diverse}}$ ↑	Word Length↑	Intra-Repetition↓
LLaVA-1.5+GPT-3.5 (Zero-Shot)	52.3	84.0	88.0	42.6	37.3	34.3
LLaVA-1.5+GPT-3.5 (Few-Shot)	52.6	84.3	88.4	46.9	50.7	23.2
Video-ChatGPT (Maaz et al., 2023a)	52.1	72.7	87.7	39.2	17.4	44.2
Video-LLaVA (Lin et al., 2023a)	52.3	73.0	87.6	32.3	28.6	34.6
VTimeLLM (Huang et al., 2023)	52.1	82.9	87.4	40.2	15.3	41.6
VTimeLLM (w/ finetune)	52.4	84.1	88.0	46.8	70.2	25.7
VideoNarrator (Ours)	52.8	**85.1**	**88.6**	**50.2**	**98.1**	10.8
a. w/o LLM LoRa finetune	52.0	83.6	86.1	42.7	42.3	**7.1**
b. w/o Storyline	52.3	84.2	88.3	49.2	96.3	8.6
c. w/o Video Clip Position	52.6	84.5	88.2	48.4	96.0	11.7
d. w/o Visual Compression	**52.9**	84.8	87.9	50.0	98.0	11.3
e. w/o Visual Memory	52.3	84.4	87.0	43.8	92.3	11.8

easier for the LLM to target relevant information. In the underlined instruction, we emphasize that the model should primarily use the information provided in the prompt and avoid using unrelated knowledge stored in the LLM.

4.3 Training

As shown in Figure 6, we construct our dataset as multimodal instruction training samples. For a video consisting of n clips, we maximize the probability of generating each script label l_i and narration s_i as follows:

$$\sum_{i=1}^{n} P(l_i, s_i | X_{\text{Instrution}}, \mathcal{K}, r_i, v_{j \leq i})$$

We update the parameters of the video projector and the video clip positional embedding layer. Additionally, we apply LoRa (Hu et al., 2022) adapter to fine-tune the LLM, enabling it to possess the ability of controllable generation and relevant knowledge combination.

5 Experiments

5.1 Experiment Settings

Implementation Details. In our experiments, we use pre-trained models as visual feature extractors as described in the above Section. In experiments on E-SyncVidStory (advertising domain), we employ the Baichuan-7B model (Yang et al., 2023) as the LLM model, known for its excellent performance in Chinese-related tasks. We construct our dataset as illustrated in Figure 6, resulting in 41k instruction samples. We apply 90% of them as the training set and the left as the testing set. Our proposed framework is trained for 15 epochs, using a learning rate of $1e^{-4}$ and a batch size of 32. The LoRA parameters are set to $r = 64$ and $alpha = 16$. The training of our 7B model took around 48 hours on 4 A6000 GPUs.

In experiments on Video Storytelling dataset (general domain) (Li et al., 2019), the LLM model is replaced by LLaMA-2 (Touvron et al., 2023) for

Table 4: Automatic evaluation results with predefined storyline. w.r.t CIDEr (C), EMScore (EMS), Info$_{\text{Sim}}$ (Sim), Info$_{\text{Diverse}}$ (Div), Word Length (Len), Script Label (Label), Intra-Repetition (IR).

	Text	Visual	Knowledge		Controllable		Fluency
	C	EMS$_{\text{ref}}$	Sim	Div	Len	Label	IR
VideoNarrator	**40.1**	**86.1**	**88.2**	**53.1**	**98.8**	95.4	9.8
a. w/o LoRa	14.9	84.5	87.1	46.0	41.0	87.2	**8.0**
c. w/o Pos.	32.2	85.1	87.6	48.1	96.5	**95.9**	8.1
d. w/o Comp.	36.7	85.6	87.9	51.8	97.8	94.1	9.2
e. w/o Mem.	29.5	85.3	87.6	49.0	92.3	93.1	8.3

Table 5: Automatic evaluation results on Video Storytelling (Li et al., 2019) dataset.

	Model Type	METEOR	CIDEr	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Length Acc.
GVMF (Lu and Wu, 2022)	Retrieval	20.7	107.7	70.5	44.3	26.9	15.9	-
VerbalizeVideos-GPT3.5 (Bhattacharya et al., 2023)	Generative	24.8	102.4	63.8	56.4	47.2	38.6	-
VerbalizeVideos-BLIP2 (Bhattacharya et al., 2023)	Generative	21.7	108.9	55.2	48.5	40.7	33.8	-
VideoNarrator (Ours)	Generative	**28.8**	**116.8**	**86.3**	**69.7**	**54.4**	**42.9**	94.1

English generation. We make minor adjustments to fit the task setting of Synchronized Video Storytelling. Specifically, we apply the text length of the ground truth as the length requirement, and use the video topic as its relevant knowledge. The training prompt is similar to that presented in Figure 6, while the system role is changed to "You are an AI assistant that can understand videos", the product information is replaced with the event topic, and the requirement for script labels has been removed.

Compared Baselines We compare our model with the following baselines. (1) *Multi-model Pipieline.* Specifically, we apply LLaVA (Liu et al., 2023) to get textual summaries for video clips and prompting GPT-3.5 for zero-shot and few-shot generation. (2) *End2end MLLMs*, including zero-shot performance of Video-ChatGPT (Maaz et al., 2023a), Video-LLaVA (Lin et al., 2023a), and VTimeLLM (Huang et al., 2023) with 7B LLMs. These models take the video as inputs and generate narrations in an end2end manner. (3) *Fine-tuned End2end MLLMs*. Specifically, we finetune the VTimeLLM model, which is effective for dense video comprehension and supports Chinese generation. More details about these baselines can be found in our Appendix A.1.

As for experiments on the Video Storytelling dataset (Li et al., 2019), we compare our model with state-of-the-art methods: the retrieval-based GVMF (Lu and Wu, 2022) and the generative-based VerbalizeVideos (Bhattacharya et al., 2023).

Ablation Study Setting. We carry out ablation studies to validate the contribution of different components in our proposed framework VideoNarrator.
a. ***w/o LLM LoRa Finetune*** Keep the LLM frozen during training in order to verify whether the LLM is effective for controllable generation.
b. ***w/o Storyline*** To verify whether the model can generate a nicely structured video story without provided explicit script labels.
c. ***w/o Video Clip Position*** To verify the effectiveness of relative clip position embedding.
d. ***w/o Visual Compression*** Retain all frames to verify the effectiveness of visual compression.
e. ***w/o Visual Memory*** Retain all video features of previous clips to verify the effectiveness of memory consolidation.

5.2 Automatic Evaluation Results

Advertising Video Storytelling. VideoNarrator can either generate a storyline to guide the generation, or use a predefined storyline to generate a story based on user preference. We conduct experiments for both situations on E-SyncVidStory.
With Generated Storyline. As shown in Table 2, our model outperforms the baseline models, even when fine-tuned on our dataset. For detailed evaluations shown in Table 3, existing MLLMs demonstrate their ability to generate visually relevant and informative stories. However, they tend to focus on a limited amount of prior knowledge and may repeat it, resulting in a lower $\text{Info}_{\text{Diverse}}$ score. Furthermore, their fluency and ability to control the length of word sequences are poor. Based on the ablation studies, we find that by fine-tuning the LLM with our instruction data, our model can achieve the ability to control the word sequence length and comprehend video features, showing an increase in controllable accuracy and visual relevance. Applying visual compression and visual memory consolidation can improve the model's ability to extract relevant knowledge and retain story coherence, resulting in better performance in knowledge relevance and fluency. Explicit storylines and relative positional embedding can encourage narrations to follow a more logical outline of persuasion, improving the performance in knowledge relevance.
With Predefined Storyline. Our model is also capable of generating stories with predefined storylines. Users can modify the generated storyline or offer a specific storyline to control the generation process, generating customized narrations. Table 4 shows the results generated with the predefined storyline, demonstrating ideal performance in controllable accuracy and other aspects.

General Domain Video Storytelling To validate the effectiveness of our proposed VideoNarrator, we conduct experiments on the Video Storytelling

Figure 7: Examples generated by our proposed VideoNarrator.

Table 6: Human evaluation results on three aspects. Geometric Mean (GM) demonstrates the overall performance.

	Relevance	Attractiveness	Coherence	GM
LLaVA1.5+GPT3.5	1.82	1.17	1.22	1.37
GPT4V+GPT4	**2.48**	2.05	2.18	2.22
VideoNarrator	2.34	**2.32**	**2.72**	**2.45**

dataset (Li et al., 2019), which covers a more general domain. As shown in Table 5, our model surpasses the state-of-the-art results. This showcases its effectiveness in synchronized video storytelling across a broader domain.

5.3 Human Evaluation and Qualitative Cases

We conduct human evaluations as well on three metrics: (1) *Visual Relevance* measures the relationship between the generated narrations and the video shots. (2) *Attractiveness* measures how well the stories can invoke the users' interests. (3) *Coherence* measures the inter-sentence coherence and the completeness of the whole story. All metrics are rated from 0 to 3. We compare our models with the few-shot generation results of existing powerful LLMs. Specifically, we apply LLaVA1.5 or GPT-4V to generate descriptions for each visual scene. The captions of LLaVA1.5 are used as inputs for GPT3.5 to generate stories, and the ones generated by GPT-4V are provided for GPT4. The prompt is shown in Appendix A.7. We randomly sample 30 stories and ask 13 advertisers to conduct the evaluation. As shown in Table 6, our model outperforms GPT-3.5. We find that GPT3.5 often generates redundant statements and sometimes describes the visual scene instead of generating attractive narration. Regarding GPT-4V+GPT4, its performance is significantly improved compared to GPT3.5. However, our model still achieves better results in terms of story coherence and attractiveness, as well as very similar visual relevance scores. Additionally, it lacks length controllability, with an accuracy of 55.0%, which impacts its practical use.

Figure 7 displays some example stories generated by our proposed framework.

6 Conclusion

This paper introduces the new task of synchronized video storytelling, which aims to generate synchronous narrations for sequential video scenes. These narrations are guided by a structured storyline and can incorporate relevant knowledge, resulting in a coherent and informative story. The new task is more practical than existing tasks of dense video captioning and visual storytelling because the generated narrations can be directly combined with the visual scenes, creating an engaging video-format story. We collect a benchmark dataset called E-SyncVidStory and introduce an effective framework named VideoNarrator. Both automatic and human evaluations verify the effectiveness of our proposed framework.

Limitations

In this work, our main focus is on generating synchronous narrations based on given videos. However, in real life, it is also important to transform unstructured video clips into well-structured ones. This aspect can be further explored in future research with the support of our proposed E-SyncVidStory. Additionally, our VideoNarrator model still faces the challenge of hallucination. For instance, most training samples show the application of face cream to the face after rubbing it on the hand. As a result, the model sometimes assumes that people will do the same when viewing someone rubbing the product on their hands. To address this issue, more constraints will be incorporated in future work.

Ethics Statement

We acknowledge the Code of Ethics and Professional Conduct and strictly adhere to the rules throughout this research. There are two potential ethical issues with our work. The first pertains to the data source, while the second relates to the use of crowdsourcing services.

Data Source. The datasets are collected from public advertisement videos on the e-commerce website [5]. Given the copyright considerations, we will only release the URLs and features of the videos. Furthermore, our data source does not contain any information that identifies individuals by name or any offensive content.

Crowdsourcing Services. After refining the data using the powerful GPT-4, we only need to conduct further checking of the annotations. We have hired 5 workers to review and correct any remaining errors. Each video took approximately 2 minutes to complete, and the workers received reasonable payment for the local area.

Acknowledgements

We thank all reviewers for their insightful comments and suggestions. This work was partially supported by the National Natural Science Foundation of China (No. 62072462) and the Beijing Natural Science Foundation (No. L233008).

[5] https://www.taobao.com/

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Aanisha Bhattacharya, Yaman K Singla, Balaji Krishnamurthy, Rajiv Ratn Shah, and Changyou Chen. 2023. A video is worth 4096 tokens: Verbalize story videos to understand them in zero shot. *arXiv preprint arXiv:2305.09758*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Spandana Gella, Mike Lewis, and Marcus Rohrbach. 2018. A dataset for telling the stories of social media videos. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 968–974.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. 2023. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*.

Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-a-story: Storytelling with retrieval-augmented video generation. *arXiv preprint arXiv:2307.06940*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2023. Vtimellm: Empower llm to grasp video moments. *arXiv preprint arXiv:2311.18445*.

Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. 2020. Multimodal pretraining for dense video captioning. In *AACL-IJCNLP 2020*.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Yaman Kumar, Rajat Jha, Arunim Gupta, Milan Aggarwal, Aditya Garg, Tushar Malyan, Ayush Bhardwaj, Rajiv Ratn Shah, Balaji Krishnamurthy, and Changyou Chen. 2023. Persuasion strategies in advertisements. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 57–66.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023b. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. 2023a. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Kevin Lin, Faisal Ahmed, Linjie Li, Chung-Ching Lin, Ehsan Azarnasab, Zhengyuan Yang, Jianfeng Wang, Lin Liang, Zicheng Liu, Yumao Lu, et al. 2023b. Mm-vid: Advancing video understanding with gpt-4v (ision). *arXiv preprint arXiv:2310.19773*.

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

Youwei Lu and Xiaoyu Wu. 2022. Video storytelling based on gated video memorability filtering. *Electronics Letters*, 58(15):576–578.

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*.

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

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

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

OpenGVLab. Ask-anything. https://github.com/OpenGVLab/Ask-Anything/tree/long_video_support. Accessed February, 2023.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*.

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

Junyi Sun. 2019. jieba. https://github.com/fxsjy/jieba.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2023. Generative pretraining in multimodality. *arXiv preprint arXiv:2307.05222*.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. 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.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging

video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022a. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. 2022b. Doc: Improving long story coherence with detailed outline control. *arXiv preprint arXiv:2212.10077*.

Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7378–7385.

Dahua Lin Yue Zhao, Yuanjun Xiong. 2019. Mmaction. https://github.com/open-mmlab/mmaction.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Hui Zhang, Tian Yuan, Junkun Chen, Xintong Li, Renjie Zheng, Yuxin Huang, Xiaojie Chen, Enlei Gong, Zeyu Chen, Xiaoguang Hu, et al. 2022. Paddlespeech: An easy-to-use all-in-one speech toolkit. *arXiv preprint arXiv:2205.12007*.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018a. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Luowei Zhou, Chenliang Xu, and Jason Corso. 2018b. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

A Appendix

A.1 Details of Compared Baselines

Multi-model Pipelines Following Zhang et al. (2023), we convert the visual information of sequential video clips into textual summaries. This information, along with the product details and length requirements, is fed into a strong generative model GPT-3.5 to produce coherent narrations. Specifically, we apply key frames from each video clip (as described in Section 4.1), generate English captions, and translate them into Chinese. We combine these captions to create the visual summaries.

你是一位广告商。请结合商品信息，为视频创作具有吸引力的广告词，吸引消费者购买。你需要判断每个镜头的剧本类型，确保广告词围绕镜头的剧本类型展开，与时间间隔内的画面内容紧密相关，符合字数范围，并且与前面的广告词保持连贯。

商品信息:
{所有相关信息}

Query:
从<s1>到<e1>，字数范围为 xx-xx 个字，请生成广告词。
Answer:
好的，剧本类型为：{镜头1的剧本类型}，广告词为：{镜头1的广告文案}
Query:
从<s2>到<e2>，字数范围为 xx-xx 个字，请生成广告词。
Answer:
好的，剧本类型为：{镜头2的剧本类型}，广告词为：{镜头2的广告文案}
...

You are an expert in writing advertising. Please use the provided product information to craft appealing advertising copy for the video. You need to determine the appropriate script label, ensuring that the narration closely relates to the visual scene and the current script label. Additionally, it should remain coherent with the previously generated text.

Product Information:
{The knowledge points}

Query:
From <s1> to <e1>, word count requirement is xx-xx words, please generate the ad copy.
Answer:
Okay, the script label is: {script label 1} the ad copy is: {Narration 1}
Query:
From <s2> to <e2>, word count requirement is xx-xx words, please generate the ad copy.
Answer:
Okay, the script label is: {script label 2} the ad copy is: {Narration 2}
...

Figure 8: This figure illustrates the prompt template for VTimeLLM, with English translations provided for easy reading.

Figure 13 shows our generation prompt. For few-shot generation, we select three example video stories with the highest click-through rate from the same domain and apply them in the prompt.

End2end MLLMs Most existing MLLMs, although promising, struggle to directly generate captions aligned correctly with sequential video clips. When an entire video is inputted, they often fail to properly ground the correct time ranges for each clip (Huang et al., 2023). For such models, we generate narrations sequentially. Specifically, we input the current video clip along with the previously generated narrations to create the current narration. We make experiments with Video-LLaVA (Lin et al., 2023a) and Video-ChatGPT (Maaz et al., 2023a), which show SOTA results on video understanding tasks. Since these models are pre-trained in English only, we apply translations using the GPT-4 model. We then manually create well-designed instructions: "You are a creative advertiser who is very familiar with the advertisement video.$\backslash n$ Product Information: {knowledge points}$\backslash n$ Previous advertisement copy: {generated narrations} $\backslash n$ Please proceed with the previous advertisement copy and create a new advertisement copy inspired by the video, using between {x1} to {x2} words. Ensure that the advertisement copy closely relates to the visual scene.$\backslash n$ Advertisement copy: "

We also conduct experiments using VTimeLLM (Huang et al., 2023), a model pre-trained for time-

Social Proof/Influence	To recommend a product, confidently endorse it on behalf of oneself or a group. Provide evidence of positive expectations and experiences to support the endorsement.
Sore Point	To identify the problems, needs, or challenges that people faced before using the product.
Call to Action	To induce or encourage consumers to make a purchase, there needs to be a strong guide that describes the benefits, prices, activities, bonuses, etc.
Design of Appearance	To describe the appearance of the product, provide details such as its shape, size, color, pattern, and any notable features of the item itself or its packaging. When describing clothing products, include information about the style design, such as the cut, texture, embroidery, etc.
Ingredient/Material/Texture	To describe the texture, ingredients, or materials of the product.
Product Trial	To provide a detailed description of the process of using or experiencing the product, creating an immersive experience for customers to understand the actual usage process.
Product Effect	To describe the function and impact of the product, as well as the changes or experiences that users can expect after using it.
Specific Characteristics	To describe the special characteristics or quality of the product, including durability, water resistance, wrinkle resistance, etc.
Product Security	To display the safety certifications and tests that the product has passed.
Authoritative Certificate	To display the quality appraisal, certificate of honor or authority of the product.
Production Process	To describe the factory manufacturing process for goods.
Others	To tell other advantages, highlights and selling points of the product. Such as the Brand history.

Figure 9: The definition of the script labels.

aligned video understanding and is effective for dense video captioning. It supports Chinese generation with the ChatGLM3 (Du et al., 2022) structure. The generation prompt is shown in Figure 8.

Finegrained End2end Large Multimodal Models To the best of our knowledge, VTimeLLM is the most effective Chinese MLLM for sequential video comprehension. We fine-tune VTimeLLM on our proposed dataset, which is transformed into its QA format, as illustrated in Figure 8. As illustrated in Table 2 and Table 3, our model surpasses the fine-tuned results in all metrics.

A.2 Evaluation Metric for Knowledge Relevance

As described in Section 3.3, we evaluate two aspects of knowledge relevance: information similarity and information diversity. To evaluate Info_{Sim}, we calculate the similarity of each knowledge point with the entire sentence s_i, selecting the maximum score as the coarse knowledge similarity We also calculate the maximum similarity of each word in s_i, and the average value is considered as the fine-grained knowledge similarity. Info_{Sim} is calculated as the average of coarse and fine-grained similarity. To evaluate $\text{Info}_{\text{Diverse}}$, we check if a knowledge point k_i has a similarity higher than 0.9 with a sentence or its segmented words. If this condition is met, we consider that the sentence includes knowledge point k_i. The $\text{Info}_{\text{Diverse}}$ score is calculated as the number of covered knowledge points divided by the video time duration. Given a story $S = \{s_i\}_{i=1}^{n}$ and its prior knowledge $\mathcal{K}$, the knowledge relevance scores are evaluated by:

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

$$\text{Info}_{\text{Diverse}} = \frac{1}{T}\left|\bigcup_{si}\left\{k_t\in\mathcal{K}\middle|\max_{w\in W(s_i)\cup s_i} f_{k_t}^T f_w > 0.9\right\}\right|,$$

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

A.3 Human Evaluation Annotators

All the annotators are employees from the advertising company and have a thorough understanding of advertisement videos. The team of 13 annotators includes 5 women and 8 men, ranging in age from 22 to 35. The Pearson Correlation score of their evaluation results is 0.55.

A.4 Script Label Definition

Script Labels are predefined structures to control the generated story. In this work, we focus on advertising stories in the e-commercial era. We analyze the collected advertising stories and categorize the script labels into 12 types: social proof/influence, sore point, call to action, design of appearance, ingredient/material/texture statement, product trial, product effect, product security, specific characteristics, authoritative certificate, production process, and others. Each of them corresponds to one of the advertising persuasion strategies (Kumar et al.,

作为广告视频语音识别结果的修正专员，你的主要任务是修正商品广告视频的语音识别结果。
你需要运用你的专业知识，包括对广告的理解以及自动语音识别常见错误的了解。广告语句通常与视频画面或商品信息相关，我们会提供参考性的商品信息来帮助你理解和修正广告的语音识别结果。你的目标是确保修正后的句子语言通顺、逻辑连贯、断句准确，并保持整体连贯。

你首要的任务是纠正语音识别结果中的错误，这些错误通常导致语言表达不顺畅。你需要发现可能的错误，并进行修正。具体地，你需要根据以下三种情况进行修正：
1. 如果句子整体没有任何实际语义，你需要根据商品信息和上下文重新编写句子。但请注意，只有在视频没有口播、只有背景音乐时，才可能出现这种情况，数量不多。尽量避免对句子进行重写。
2. 如果句子明确描述的是无关的商品，你需要利用商品信息或你的专业知识，改写句子。请只对描述无关商品的部分进行改写，其余部分则进行错误修正。在进行改写时，你需要确保句子与上下文的连贯性。
3. 除了以上两种特殊情况，你需要修正句子中的错误，确保修正后的句子语言通顺、逻辑连贯、断句准确。修正过程中，可以参考商品信息和专业知识，但不能引用上下文的内容。对于专有名词，如商品成分，需特别注意，避免被误识别为发音相似的其他词。

在完成上述任务后，你需要在尽可能保持原句意思的基础上，适当优化句子，确保其流畅、通顺，与前文连接自然。

As an expert in refining automatic speech recognition (ASR) results, your primary task is to correct errors within the speech recognition results of advertisement videos.
You will need to apply your professional knowledge, including an understanding of advertisements and common errors in automatic speech recognition. Advertising narrations are usually related to the video scenes or product information, and we will provide information to help you understand and correct the speech recognition results of the advertisements. Your goal is to ensure that the corrected sentences are fluent, logically reasonable, punctuated accurately, and maintain overall coherence.

Your first task is to correct errors in the speech recognition results. These errors often lead to confusing language expressions. You need to identify and analyze potential errors and make corrections. Specifically, you need to make revisions according to the following three scenarios:
1. If the sentence as a whole has no actual meaning, you need to rewrite the sentence based on the product information and context. However, please note that this situation may only occur when there is no voiceover in the video, only background music, and it is not common. Try to avoid rewriting sentences as much as possible.
2. If the sentence clearly describes an unrelated product, you need to use the product information or your expertise to rewrite the sentence. Please rewrite only the part that describes the unrelated product, while correcting errors in the rest of the sentence. When rewriting, you need to ensure the sentence's coherence with the context.
3. Besides the two special scenarios mentioned above, you need to correct mistakes in the sentence to ensure that the corrected sentence is fluent, logically coherent, and punctuated accurately. During the correction process, you may refer to the product information and professional knowledge, but you cannot quote the context. Special attention should be paid to proper nouns, such as product ingredients, to avoid misrecognition as other words that sound similar.

After completing the above tasks, you should optimize the sentences appropriately to ensure they are smooth, fluent, and naturally connected to the preceding text, while maintaining the original meaning as much as possible.

Figure 10: The prompt for automatic ASR refinement, with English Translation for easy reading.

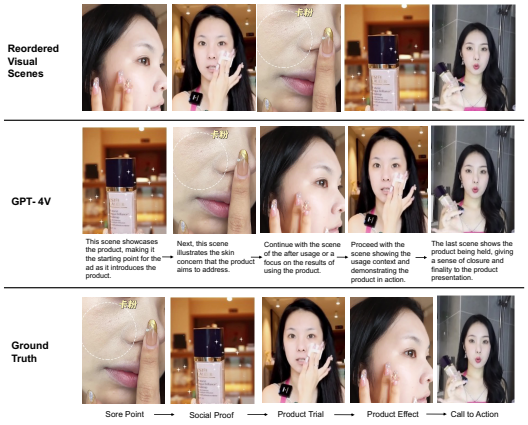

Figure 11: This figure illustrates the sorted results by GPT-4V.

2023). The definition of the 12 script labels is displayed in Figure 9.

A.5 Discussions of Further Research

In our paper, the visual scenes are logically sorted. But at times, a user may wish to input unordered visual recordings and obtain sorted clips with a coherent narrative. We also hope to invoke additional research on the reordering task using the annotation provided in our proposed dataset. In Figure 11, we present the quantitative results sorted by the robust GPT-4V model. These results are still imperfect. Considering its reasoning process, the usage process should be positioned between the product's pre-use and post-use stages. Starting the advertisement with the pain point could make it more appealing, capturing the users' needs right from the beginning. This highlights the need for further research.

A.6 Prompt for Data Processing

The prompt for ASR refinement is displayed in Figure 10. The prompt for script label classification is displayed in Figure 12.

A.7 Prompt for Few-Shot Story Generation

The prompt for few-shot story generation with GPT3.5 and GPT4 is displayed in Figure 13.

你是一位熟悉淘宝商品和广告文案的广告商。我将为你提供商品信息和相应广告的文案。你的任务是根据文案的内容，判断文案的剧本类型。
以下是广告文案的剧本类型及其定义：

{剧本类型及定义}

请根据以上剧本类型的定义，结合商品信息对文案的内容进行分析和理解，输出文案对应的剧本类型。请注意，文案的剧本类型只能在给定的12种类型中选择。同一段文案可能包含多种剧本类型，如果有多种剧本类型，用"+"进行连接。
现在开始你的任务吧。

You are an advertiser familiar with Taobao products and advertising copywriting. I will provide you with the product information and the corresponding advertisement copy. Your task is to judge the type of script based on the content of the copy.
Below are the types of advertisement scripts and their definitions:

{Script Labels and Definitions}

Please analyze and understand the content of the copy in conjunction with the product information, based on the definitions of the script types given above, and determine the corresponding type of script for the copy. Note that the type of script for the copy can only be chosen from the given 12 types. A single piece of copy may contain multiple script types; if there are multiple types, connect them with a "+."
Now begin your task.

Figure 12: The prompt for script label classification, with English Translation for easy reading. The definition of script labels are detailed in Figure 9

你是一位广告商，擅长为商品撰写具有吸引力的广告文案，并且能够准确理解商品信息和视频信息。

我将为你提供商品信息以及该商品对应的广告视频的每个镜头的信息。你的任务是结合商品信息，为视频的每个镜头撰写符合当前字数要求的广告文案，吸引消费者购买。你需要判断每个镜头的剧本类型，确保文案围绕镜头的剧本类型展开，与当前场景紧密相关，并且与前面的文案保持连贯。

请注意，在撰写文案时，主要使用到提供的商品信息以及镜头的场景信息。

{任务样例}

输入:
商品信息:
{所有相关信息}
视频:
  片段1:
  场景: {当前场景的描述文本}
  字数要求: 22-28个字
输出:
  文案:

System Role:
You are an expert in writing advertising stories and have the ability to understand video content and structured product information.

Instruction:
I will provide you with information about the product, as well as details about each clip in the video. Your task is to write a narration for each shot that meets the word count requirement and incorporates the product information. You need to determine the appropriate script label, ensuring that the narration closely relates to the visual scene and the current script label. Additionally, it should remain coherent with the previously generated text. Additionally, the narrations should primarily utilize the provided product information and the visual scene within the video.
Now please start working.

{Few-Shot Examples}

User:
Product Information:
{The knowledge points}
Video:
  Video Clip 1:
  Visual Scene: {Visual Frame Descriptions, generated by LLaVa or GPT-4V}
  Word count requirement: 22-28 words
Output:
  Narration:

Figure 13: The prompt for few-shot generation, with English Translation for easy reading.