

# Shot2Story20K: A New Benchmark for Comprehensive Understanding of Multi-shot Videos

Mingfei Han<sup>2,1,3</sup> Xiaojun Chang<sup>2,4</sup> Heng Wang<sup>1</sup> Linjie Yang<sup>1</sup>  
<sup>1</sup>Bytedance <sup>2</sup>ReLER Lab, AAIL, UTS <sup>3</sup>Data61, CSIRO <sup>4</sup>MBZUAI  
<https://mingfei.info/shot2story/>

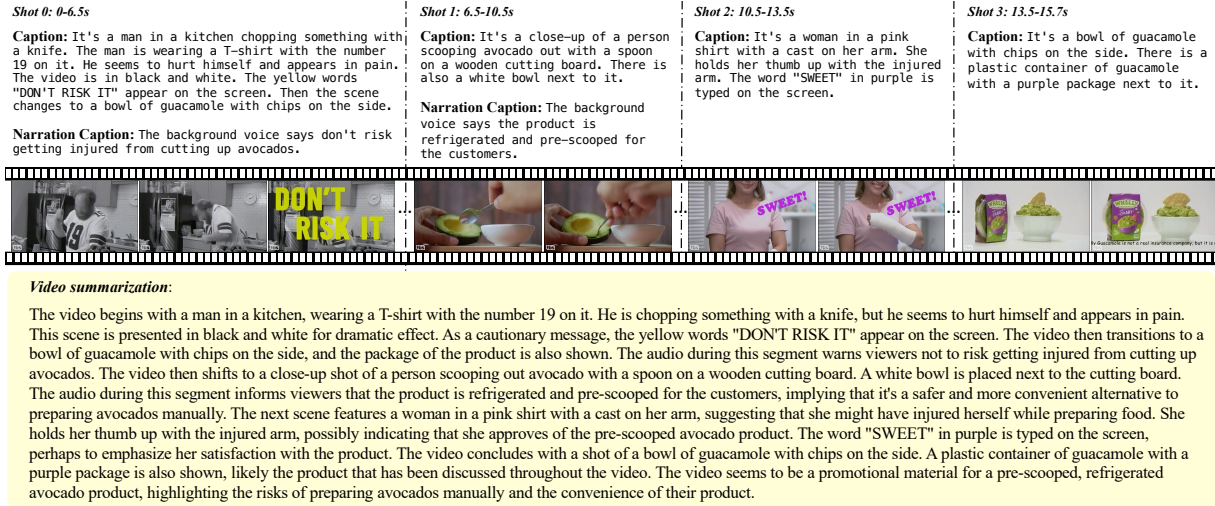


Figure 1. An annotated example of our Shot2Story20K with sing-shot visual captions and narration captions. Moreover, we provide coherent and reasonable video summaries to facilitate comprehensive understanding of multi-shot videos.

## Abstract

A short clip of video may contain progression of multiple events and an interesting story line. A human need to capture both the event in every shot and associate them together to understand the story behind it. In this work, we present a new multi-shot video understanding benchmark Shot2Story20K with detailed shot-level captions and comprehensive video summaries. To facilitate better semantic understanding of videos, we provide captions for both visual signals and human narrations. We design several distinct tasks including single-shot video and narration captioning, multi-shot video summarization, and video retrieval with shot descriptions. Preliminary experiments show some challenges to generate a long and comprehensive video summary. Nevertheless, the generated imperfect summaries can already significantly boost the performance of existing video understanding tasks such as video question-answering, promoting an under-explored setting of video understanding with detailed summaries.

## 1. Introduction

Video captioning is a long-standing video understanding task to facilitate open-world video analysis with the help of human-annotated captions. Since a video may contains multiple events, dense captioning benchmarks (Ego4D [10], YouCook2 [36], ActivityNetCaps [13]) are tailored to capture the information of multiple events in a video ranging from 3-20 minutes. However, even within seconds, we find that there are already more than one single event in a lot of daily videos such as news broadcast, tutorial videos, and movies. Specifically, shot transition, which is a common technique to transfer from one event to another, or to switch the viewpoint of a single event, happens less than every 4s for average English movies after 2010 [5]. Although some existing captioning benchmarks [13, 33, 36] already use multi-shot videos, they often annotate the captions in a coarse-grained manner, either providing a holistic caption or asking annotators to subjectively choose the boundary of each event. To better accommodate the multi-shot formation of videos, we believe a new video benchmark with rich

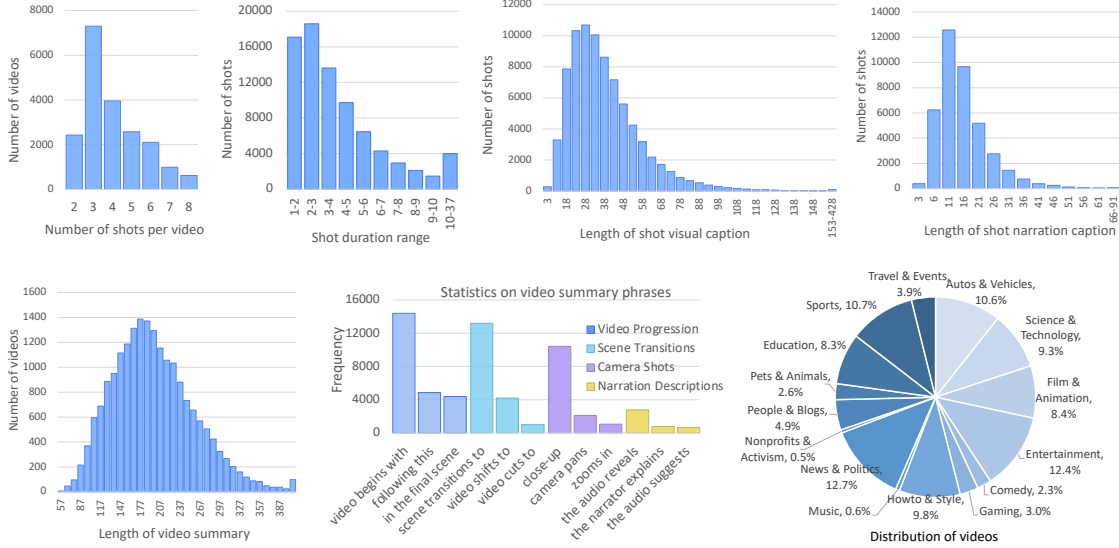


Figure 2. Statistics of Shot2Story20K . Our dataset comprises videos with 2 to 8 shots each. Most shots range from 1 to 5 seconds, accompanied by detailed visual captions and narration captions. It features extensive summaries, highlighting video progressions, transitions, camera cuts and narration descriptions, with statistics of frequent expressions depicted in the figure.

textual descriptions based on video shots is favored in the research community.

On the other hand, multi-shot videos are often accompanied by rich narrations that relates to the different events happening in the video. A model needs to capture both the visual and audio signals to understand the underlying story. Specifically, narrations may contain key information that cannot be inferred from pure visual information only. See Figure 1, without the narration, a viewer is unable to capture the relationship between the man’s action and the avocado product in the first shot.

In this work, we propose a new benchmark Shot2Story20K for audio-visual understanding of multi-shot videos. We collected a dataset of 20,023 short videos where the average number of shots in each video is 4.0. For each video shot, we annotate a detailed textual description for the video frames and another textual description for the human speech. We also leverage a state-of-the-art large language model (LLM) GPT-4 [23] to generate a long textual video summary from the annotated clip descriptions, which are further verified by human annotators. The summary includes additional details such as transitions of different shots, progression of multiple events, and mapping of the subject identities in different scenes. An overview of our dataset can be seen in Figure 2.

To benchmark the advances of multi-modal video understanding, we designed several distinctive tasks using our dataset, including single-shot captioning, multi-shot summarization, and video retrieval with shot description. We design and implemented several baseline models using a frozen vision encoder and an LLM, by prompting the LLM

with frame tokens and ASR (Automatic Speech Recognition) text. Through extensive experiments, we show that: (1) the ASR text is critical to understand the complex multi-shot scenario, (2) processing the video as a whole without the shot-structure degenerates the model’s capacity of understanding the multi-shot video, (3) the summarization model trained on our benchmark can be generalized to other datasets with longer durations (ActivityNet) and out-of-domain topics (MSRVTT). Without any bells and whistles, we attains competitive results on zero-shot video question-answering by converting the problem into pure text-based QA with the generated video summaries.

## 2. The Shot2Story20K benchmark

### 2.1. Overview

Our new benchmark Shot2Story20K contains 20,023 videos. The length of each each video is ranging from 10s to 40s. For each video, we first use a off-the-shelf shot detection method TransNetV2 [28] to split it into shots. For each video shot, we annotate captions for both visual and audio information. Then we further annotate video summaries based on the annotated shot captions. Figure 2 shows an overview of our dataset with some key statistics. An example of one annotated video is shown in Figure 1.

### 2.2. Data preparation

We source videos for our dataset from the public video benchmark HDvila100M [34]. It offers a large collection of narrative videos, comprising 3M YouTube videos segmented into 100M clips, each about 13 seconds long. We

Dataset	Annotation	Multi-shot Video	Multi-event Descriptions	Audio Captions	Detailed Summary	Summary Length	#Videos	Avg. Duration
MSRVTT [33]	M	✓	✗	✗	✗	-	10K	15s
ActivityNet Caps [13]	M	✓	✓	✗	✗	52.4	20K	3min
VideoStorytelling [15]	M	✓	✓	✗	✓	162.6	105	12.5min
Ego4D [10]	M	✗	✓	✗	✗	-	10K	23min
YouCook2 [36]	M	✓	✓	✗	✗	67.8	2K	6min
VAST [3]	G	✓	✗	✓	✓	32.4	27M	5~30s
Shot2Story20K	M+G	✓	✓	✓	✓	201.8	20K	16.7s

Table 1. High level comparison of our dataset to previous ones. The summary length of ActivityNet and YouCook2 are their combined length of captions in one video. M and G stands for manual and generated, respectively.

choose this data source for its concise yet complex multi-shot formats, diverse topics, and abundant ASR content. Since we prefer videos with both rich visual and ASR information, we design several filtering techniques to exclude those videos with either low visual-ASR correlation or static visual content.

We start with keeping video clips with durations between 10 to 40 seconds, since we observe that the majority of the video clips from HDvila100M fall in this range. Then we remove videos with more than 8 shots due to the heavy annotation cost. We also notice that the video segments with too many shots in HDvila100M tend to be slideshows or image collages that deviates from our focuses. Further, to harvest videos with rich visual-ASR correlations, we set up a metric between video shots and ASR texts. Specifically, we uniformly sample 4 video frames for each shot and obtain the cosine similarity score between the video shot embedding and the text embedding using CLIP [25] encoders. We only keep the videos containing at least one shot that is visually correlated to ASR with a threshold of 0.25. In the next step, in order to obtain videos with diverse shot contents, we set up an inter-shot metric to filter out the videos with similar adjacent shots. We compute the cosine similarities between embeddings of adjacent shots and keep the videos with all inter-shot similarity scores smaller than 0.9. Finally, to further remove the videos with static contents, we adopt an intensity-based scene changes filter in PySceneDetect<sup>1</sup> with a threshold 11 in our segmented shots. If the filter is unable to detect new segments at a low threshold, it is conceivable that the shot contains static contents. We only keep the video clips in which all shots contain no static content based on our filtering method.

As a result, from a total of 1.1M sampled video clips from HDvila100M, we obtain 20,023 video clips that meet our quality standard. The number of shots in each video is from 2 to 8. These videos are then shared with our annotators for further annotations.

### 2.3. Annotation of single-shot captions

After using TransNetV2 to divide the target videos into video shots, we ask annotators to annotate both visual-only captions and audio-related captions for each shot. We split the annotation of these two captions to facilitate separate modeling of these two types of information source.

For visual-only caption, we require annotators to describe the major subjects and events in the video. Since it is an open-world setting, the videos can be quite diverse and hard to describe. In order to reduce the difficulties of annotating a caption from scratch, we generate an initial video caption using MiniGPT-4 [38] by sampling 4 image frames from the video clip and prompting the model using the prompt below.

*MiniGPT-4 prompt: ###Human:<Img>Frame1</Img><Img>Frame2</Img><Img>Frame3</Img><Img>Frame4</Img>Please describe this video. Do not include details that you are not sure of. For example, if there is text in the image, do not include the content of the text if they are not clearly shown. ###Assistant:*

Although MiniGPT-4 is originally designed for image understanding, empirically it is able to generate captions for videos, both comprehensively and reasonably. It is able to describe different subjects including person, animals, food, tools, and virtual objects like animated characters. We ask annotators to correct any mistakes they find in the generated captions, and to add missing details to the captions. The mistakes include incorrect description of the object categories, attributes, actions, facial expressions etc. Also, there might be some subjective description generated by MiniGPT-4 such as emotion and atmosphere. We ask annotators to remove all these subjective descriptions. An example of this annotation process is shown in Supp. Sec. 5.1, where the annotator corrects the caption from “standing in front of the car” to “getting close to the car”, and adding a missing detail of “a close-up shot of the front”. In this way, we find the annotation speed significantly faster ( $\sim 3\times$ ) compared to writing a caption from scratch. On the other hand, we find the captions generated this way has more co-

<sup>1</sup><https://www.scene detect.com/>

herent style and tend to cover more details of the video.

In contrast to the traditional video captioning benchmarks [13, 33, 36], we also annotate narration captions in addition to the visual-only captions. Different from existing audio captioning benchmarks [9], we focus more on human speeches rather than acoustic events. The annotators are required to associate the human speech with the video content and summarize the main idea of the speech. We require annotators to describe the source of the speech using visual information. For example, if someone is talking, the annotator needs to describe which person in the video is talking. If the human speech refers to some object in the video, the annotator is required to describe which object in the video the speaker is referring to. Note that the speaker identity and reference of visual objects are critical information for understanding a video that cannot be trivially obtained using existing algorithms. There are existing research on speaker identification [12] and visual grounding [1, 37], but they only work well on constraint scenarios. Given this annotation process, our narration captioning task requires a joint understanding of visual and audio signals.

## 2.4. Annotation of video summary

To create video summaries with the annotated video-shot captions, we leverage an LLM-based approach. Specifically, we form a text prompt with incorporating all shot captions and ASR text included, and uses GPT-4 [23] to generate a cohesive summary. The text prompt we use is shown in Supp. Sec. 5.2. The quality is assured through further review and correction by our annotators.

We prompt GPT-4 to produce coherent, fluent text summaries with transition expressions such as “the video begins”, “following this”, and “in the final scene” to connect video-shot descriptions. The generated annotations also encompass a higher-level understanding of shots, using key phrases such as “scene shifts back to” and “returns to the scene” to denote recurring scenes across different shots. Notably, GPT-4 often identifies and links the same subjects across scenes without relying on explicit re-identification models. It draws on descriptive and attributive text from our shot captions like “a newsroom” or “a man wearing a black suit” to infer scene or subject identity. To ensure quality, annotators carefully review and correct any inconsistencies in scene or subject references within these summaries.

## 2.5. Comparison to existing benchmarks

Compared to existing video description datasets, our dataset is more challenging due to the explicit modeling of the multi-shot nature of web videos. Our textual description includes both shot-level captions and video-level summaries, combining visual and audio understanding, which provides a unique test bed for multi-modal video understanding. Table 1 shows a high-level comparison of our new dataset

with existing video captioning benchmarks. Most existing video captioning benchmarks, such as MSRVT [33], YouCook2 [36] and ActivityNet Caps [13], also use multi-shot videos as annotation source, but they either annotate a holistic caption for the video (MSRVT) or ask annotators to decide the boundary of different events. In our study, we observe that video shots naturally create a sequence of related events, leading us to annotate distinct captions for each shot. Ego4D [10] only annotates dense visual captions but not audio captions for relatively long egocentric videos. Video Storytelling [15] is a small-scale dataset with annotations of multiple events in a videos and provides a summary of the video by concatenating all captions.

A recent work VAST [3] feed generated video and audio captions into an LLM to generate video summary. However, their work processes a multi-shot video as a whole and lacks the granularity of the events in different shots. Additionally, VAST directly uses predicted captions without any human verification, which indicates their video summaries can be noisy and containing biases from the captioning models. Our dataset stands out from VAST with its more detailed visual and audio shot captions. These captions, averaging 35.3 words for visual and 17.8 words for audio, are the result of a thorough manual annotation process. Although our video summary is also generated using an LLM, it is further verified by annotators to make sure there is no hallucinated details from the LLM. Our dataset has an average length of 201.8 words for the video summary, which is much longer than existing benchmarks, and longer than the combined length of captions in one video in ActivityNet and YouCook2.

# 3. Tasks and Experiments

## 3.1. Basic settings

For all the tasks described in this section, we follow the same training/validation/test split. Specifically, the number of videos for training, validation, and test set are 14016, 1982 and 4025, respectively. We resize the frames to  $224 \times 224$ . We employ ViT-G/14 from EVA-CLIP [8] and Q-Former from BLIP-2 [16] as visual encoder, and Vicuna v0-7B [4] as the language model. We load pretrained Q-Former from MiniGPT-4 [38]. In training, we update only Q-Former parameters, keeping the ViT and LLM frozen. We adopt AdamW [20] as our optimizer and use a learning rate of  $8e-5$ . We train the models for 40 epochs with a batch size of 128 for single-shot video captioning and narration captioning. We finetune our video summarization models on the single-shot captioning model with a batch size of 16.

## 3.2. Single-shot video captioning

To understand the visual content of each video shot, we introduce the single-shot video captioning task. Note that the



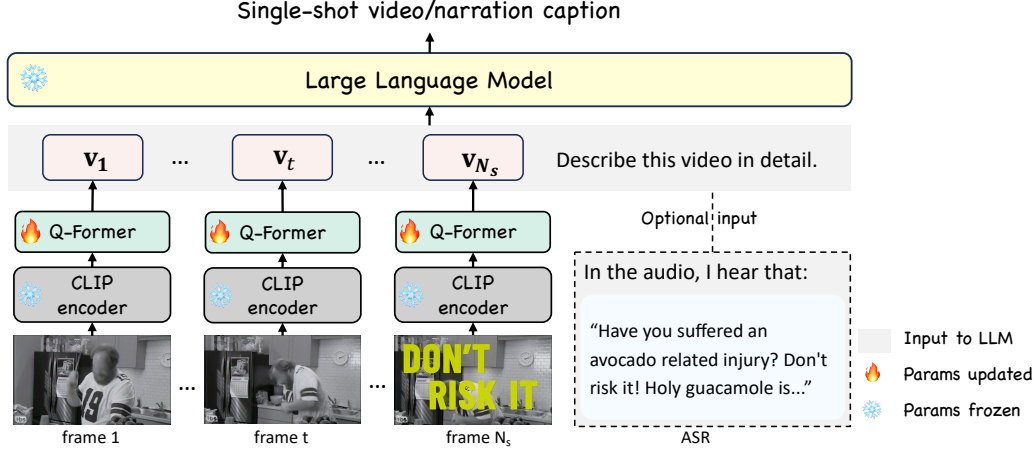


Figure 3. Model structure for video-shot captioning. Visual tokens from the CLIP [25] visual backbone and Q-Former [16, 38], along with text prompts, form the input to the LLM [4]. ASR input is optional for single-shot video captioning.

Modalities	B	M	R	C
V	10.5	16.0	30.1	38.8
V+A	10.7	16.2	29.6	37.4

Table 2. Performance of single-shot video captioning task. V and A means Visual and ASR.

task is to generate descriptions for individual video shots, while ASR information can be leveraged to improve the accuracy of the captions. For this task, we adapt the framework of MiniGPT-4 [38], with the model structure depicted in Figure 3. Specifically, we adopt the similar structure as we generate pseudo captions for data annotation in Sec. 2.3. First, we sample  $N_s$  frames from a video shot, and encode them using a fixed vision encoder, then feed the encoded features to a Q-Former to produce visual tokens. The visual tokens are appended into a text prompt and the LLM is asked to generate a caption for this video shot.

We compare two model variants on this task. One is with the ASR text as additional context cues in the text prompt and the other is without the ASR information. We evaluate our models using BLEU@4 [24] (abbreviated to B), METEOR [6] (abbreviated to M), ROUGE [19] (abbreviated to R) and CIDEr [30] (abbreviated to C), and show the results in Table 2. It shows that inclusion of ASR-derived texts yields a modest enhancement in the B and M by 0.2. Conversely, it incurs a decrement of 0.5 and 1.4 in R and C, respectively. These results imply that ASR text complements visual data without introducing discrepancies, yet posing integration challenges for augmenting single-shot video captioning performance. Figure 5 (a) displays output examples of our model’s single-shot video captioning. It accurately details visual elements within the shot, effectively capturing actions like “gesturing with her hands” and articulates secondary elements within a scene like “a doll on the couch”.

Method	Modalities	B	M	R	C
VALOR [2]	Audio	6.6	10.0	23.9	13.5
Ours	A	4.7	17.1	30.3	130.9
	V+A	18.8	24.8	39.0	168.7

Table 3. Performance of single-shot narration captioning task. V and A means visual and ASR.

### 3.3. Single-shot narration captioning

Human narration is another critical factor to understand a multi-shot video. It often provides information of the background knowledge and commentator’s view on visual events. We conduct experiments to predict the narration caption of a video-shot and name this task single-shot narration captioning. We adopt the same model structure as single-shot video captioning with the ASR text as additional input, except that the prediction target is the narration caption. We compare with existing audio captioning model VALOR [2]. We finetune VALOR on our single-shot narration captions and show the results in Table 3. We also add another baseline model that only takes ASR text as input and predicts the narration captions using Vicuna [4].

Since our narration captions contain descriptions about the related visual information as well, e.g. the subject, referred objects etc, using only ASR text does not produce satisfactory results. The baseline model VALOR is unable to capture the rich ASR text information with only the raw audio, leading to a weak performance of 13.5 in CIDEr. Our model combining visual and ASR text can generate reasonable narration captions on most cases. As shown in Figure 5 (a), our model identifies narration sources and aptly describes spoken content, as highlighted by phrases like “the background voice says” and “the man in a hat is talking”.

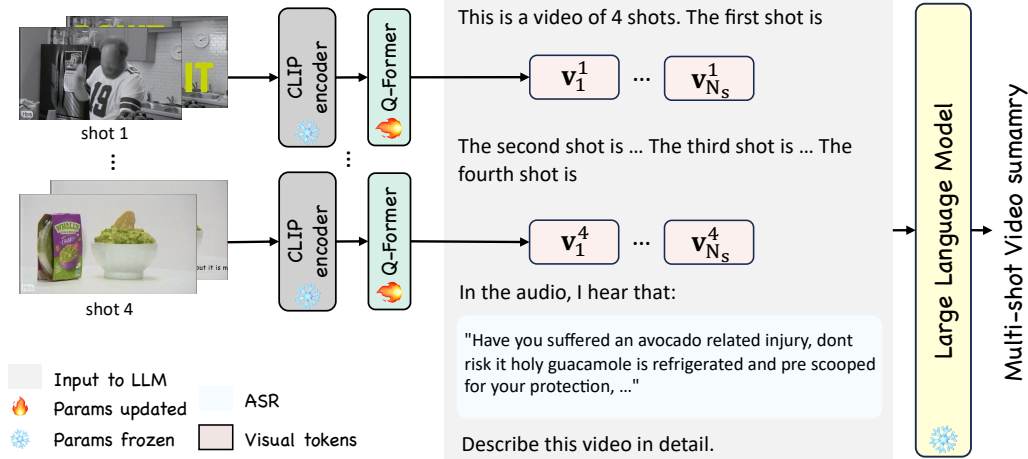


Figure 4. Model structure for multi-shot video summarization model SUM-shot. We arrange visual tokens in a multi-shot format to encapsulate multi-shot information. Additionally, ASR text is incorporated for audio-visual video summarization.

### 3.4. Multi-shot video summarization

Multi-shot video summarization is a new task that is distinct from existing video description tasks. It requires the model to understand the shot structure of the given video and to provide a coherently paragraph to describe the progression of events in the different shots. In this section, we experiment with three model variants. The first model SUM-text uses a two-stage approach, first generating captions using our video-shot captioning model for each video shot, then embed the generated captions into a text prompt as the input to the LLM (Vicuna-v0 [4]) to generate a video summary. The associated prompt can be seen in Supp. Sec. 6.2. The second model SUM-holistic uses similar model as Figure 3. We uniformly sample 16 frames from the full video clip and prompt the LLM with frame tokens and ASR text. The third model SUM-shot uses a more refined framework by sampling 4 frames in each video shot and prompting the LLM with frame tokens from different shots, as is shown in Figure 4. Compared to SUM-holistic and SUM-shot, SUM-text is not trained end-to-end and may lose critical information with the captioning step, for example, it cannot capture the correspondence of the same subject in two shots. SUM-holistic does not have the shot information explicitly and rely on the LLM to parse the video shots using the provided frame features. SUM-shot is given the shot structure as input, which makes it easier to generate descriptions based on the different shots. We compare with Video-ChatGPT [22] by instruction-tuning their model on our video summary data without the ASR input.

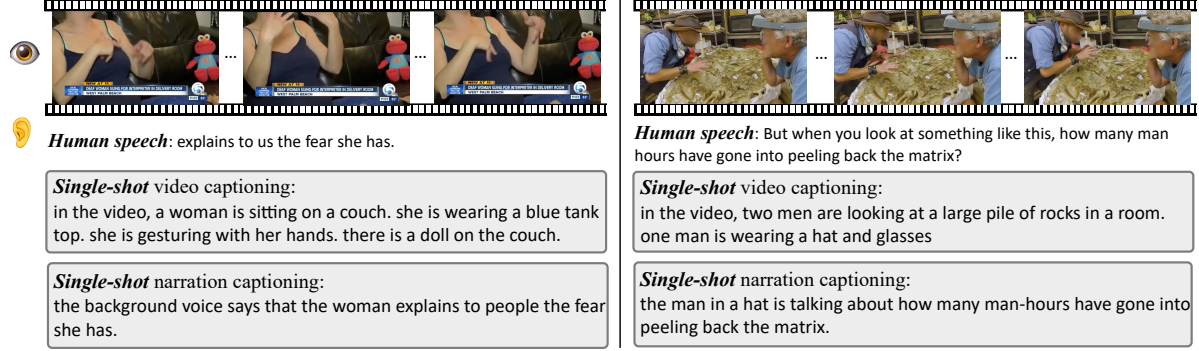
Table 4 shows the results of the three models. It is shown that SUM-text achieves the overall best performance, although it is a two-stage model with pre-generated shot captions. SUM-shot is slightly worse than SUM-text, indicating that better model design needs to be explored for end-to-end video summarization. SUM-holistic is consistently

Model	E2E	ASR	B	M	R	C
Video-ChatGPT [22]	✓	✗	5.0	14.0	19.7	1.2
SUM-shot w/o ASR	✓	✗	9.8	18.4	24.9	4.7
SUM-text	✗	✓	12.2	20.4	27.1	9.2
SUM-holistic	✓	✓	10.9	18.3	26.2	6.3
SUM-shot	✓	✓	11.7	19.7	26.8	8.6

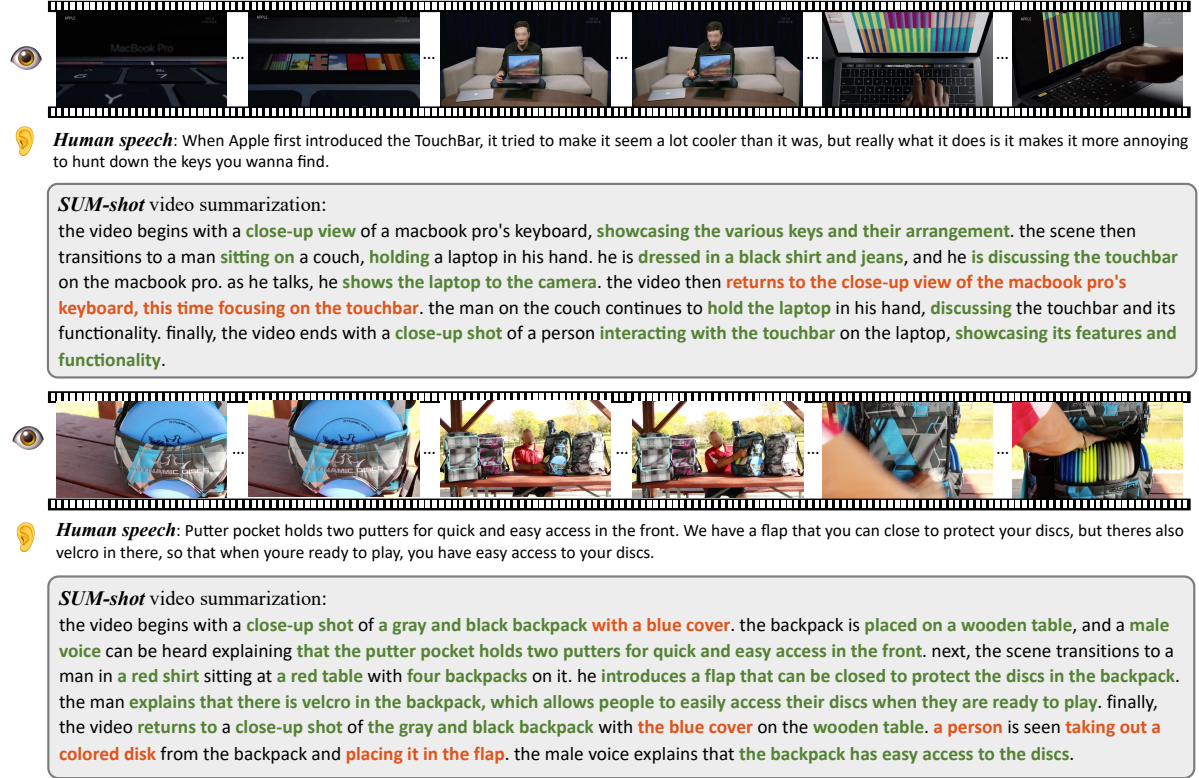
Table 4. Performance of models on video summarization. E2E means whether the model is trained in an end-to-end approach.

worse than SUM-shot, showing the importance of the shot structure in predicting a video summary matching the transition of shots. SUM-shot w/o ASR underperforms compared to SUM-shot and SUM-holistic, highlighting ASR’s significance in multi-shot understanding. Video-ChatGPT is not able to match the performance of our models, potentially due to their weakness in processing multiple scenes and lack of ASR input. Video-ChatGPT directly encodes the whole video into a sequence of tokens and may lose a lot of details in the frames, while ours directly feeding frame tokens into LLM without compressing them.

Figure 5 (b) illustrates our SUM-shot model’s predictive capabilities. The model adeptly narrates event sequences with appropriate emphasis. For instance, in the MacBook-example, it not only details the keyboard but also rationalizes the display of various keys, aligning with the ASR data about the touchbar discussion, thus crafting a coherent summary. Nonetheless, some predictions, marked in red, are erroneous, such as the non-existent “returns to a close-up view of the macbook” shot. These inaccuracies likely stem from the LLM’s tendency to “hallucinate” plausible yet non-factual details. Despite these errors, the model demonstrates a proficiency in generating consistent and nuanced summaries, highlighting both the potential of our model and the challenges that our dataset presents for future research.



(a) Example predictions of single-shot video captioning and narration captioning.



(b) Example predictions of multi-shot video summarization.

Figure 5. Example predictions of our models. (a) demonstrates our model’s single-shot video captioning, producing precise descriptions and identifying narration speakers, e.g., *gesturing with hands*, *a man in a hat speaking*. (b) shows multi-shot video summarization, with accurate captions in green and errors in red, illustrating the model’s ability to narrate event sequences and maintain subject consistency, as seen in the progression from *close-up of a backpack* to *transitions to a man* and *return to the backpack*.

### 3.5. Video question-answering with summary

Since the generated summaries are long and complex, the traditional captioning metrics (B, M, R, C) may not reflect the true quality of the generated summaries. We thus adopt another video understanding task, zero-shot video question-answering (QA), to further evaluate the quality of our generated summaries. Existing work [11] directly uses image captions as input to an LLM to generate question response. However, no such work has been done for videos.

Specifically, we directly apply our video summarization model on video QA benchmarks MSRVT-QA [32] and ActivityNet-QA [35] by splitting the testing videos into video shots and feeding them into the SUM-shot model. The generated summaries and the associated questions are then fed into a Vicuna model to derive the answers with the prompt shown in Supp. Sec. 6.3.1. Note there is no adaptation or finetuning conducted for the Vicuna model. Since the original answers in the QA benchmarks are very short and the generated responses from LLM tend to be a long

Method	Pretrain Datasets	T2V			T2S			V2T		
		R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
Alpro [14]	WebVid-2M+CC-3M	46.3	69.5	78.4	50.3	76.3	83.2	45.2	69.8	78.2
Clip4clip [21]	CLIP400M	47.2	70.4	77.6	52.4	78.2	85.4	48.9	70.2	78.1
UMT [18]	CLIP400M+UMT25M	66.3	81.8	85.8	68.6	88.4	92.0	64.9	82.3	86.2

Table 5. Comparison of performance for text-to-video (T2V), text-to-shot (T2S), and video-to-text (V2T) retrieval tasks.

sentence, we leverage the `gpt-3.5-turbo` model to generate a binary decision of whether the answer is correct, following Video-ChatGPT [22]. We compare our results with Video-ChatGPT [22], MovieChat [27] and VideoChat [17] as in Table 2. Note that Video-ChatGPT and VideoChat both use large amount of instruction tuning data to learn to directly generating answers from visual features and the text prompt, while ours bypasses instruction tuning by distilling the video information into a video summary. Additionally, for a direct comparison, we evaluate Video-ChatGPT on question-answering in the same methodology as ours. As shown in Table 6, our model outperforms Video-ChatGPT by a large margin. Our model also follows the zero-shot QA settings since the model only uses Shot2Story20K as training data. Note that MSRVTt contains a large portion of videos with out-of-domain topics such as tv shows and food, while ActivityNet has much longer videos than our training videos. This validates the robustness and transferability of our model across different topics and longer videos. This surprisingly good result indicates that a comprehensive and detailed video summary is a high quality abstraction of the video, facilitating a wide range of tasks including video QA and video-based conversation. Example summaries on MSRVTt and ActivityNet can be seen in Supp. Sec. 7.3.

### 3.6. Video retrieval with shot description

Text-based video retrieval is another task to evaluate multi-modal video representations. Traditional video retrieval often utilizes highly condensed text descriptions with benchmarks such as MSRVTt [33], LSMDC [26], and VA-TEX [31]. Retrieval models can simplify the problem by leveraging key objects / actions in the video without understand more complex details such spatial-temporal information and user intent. We present a distinct setting for retrieval with only descriptions of one video shot. Specifically, we design three settings: (1) using a shot description as query source to query the corresponding video (T2V). (2) using a shot description as query source to query the specific shot (T2S). (3) using a video as source to query a randomly sampled shot description in this video (V2T).

We report results on the testing set of our benchmark with 4025 videos, including 15913 shots. We evaluate several baseline models including Alpro [14], CLIP4clip [21], and UMT [18] and show the results in Table 5. In the three retrieval tasks, Alpro underperforms relative to Clip4clip

Model	IT	QA Input	MSRVTt QA	ActivityNet QA
VideoChat [17]	✓	V+T	45.0	26.5
Video-ChatGPT [22]	✓	V+T	49.3	35.2
MovieChat [27]	✓	V+T	49.7	<b>51.5</b>
Video-ChatGPT [22]	✓	T	53.7	37.4
SUM-shot+Vicuna	✗	T	<b>56.8</b>	47.4

Table 6. Performance on video question answering. IT means whether the model uses video-text instruction tuning data. All methods follow the zero-shot manner.

by approximately 2.0 points in R@1, while UMT outperforms Clip4clip significantly, with an R@1 improvement of 19.1 for T2V and 16.2 for T2S. The performance comparison confirms that refined video-language alignment is crucial for retrieval accuracy. While Alpro employs regional token alignment and CLIP4clip uses global video-text matching, UMT advances the field with its R@1 improvements, utilizing masked modeling and distilling a ViT [7] for more detailed alignment. In light of these findings, our Shot2Story20K, enriched with ASR information that closely aligns with visual elements, presents an opportunity to harness ASR as a natural linkage for improving video-text alignment, potentially guiding future enhancements in this domain. Additionally, a comparison between video and shot retrieval tasks reveals that T2V presents a greater challenge than T2S, aligning with our hypothesis that retrieving a full video using a shot caption necessitates a more detailed understanding of the video. It confirms the capacity of our dataset to facilitate detailed and complex video understanding tasks.

## 4. Conclusion

In this work, we present a large-scale video understanding benchmark with annotations based on video shots. We provide detailed textual descriptions for each shot as well as a comprehensive video summary for the whole video. With the rich and diverse descriptions, our benchmark serves as a playground for more powerful multi-modal video understanding models, ready to be extended for a range of other video understanding tasks, such as video question answering, visual grounding, and video-based conversation.



## References

- [1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pages 5803–5812, 2017. 4
- [2] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Wein-ing Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 5
- [3] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023. 3, 4
- [4] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, 2023. 4, 5, 6
- [5] James E Cutting, Kaitlin L Brunick, Jordan E DeLong, Catalina Iricinschi, and Ayse Candan. Quicker, faster, darker: Changes in hollywood film over 75 years. *i-Perception*, 2(6):569–576, 2011. 1
- [6] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 8
- [8] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. *arXiv preprint arXiv:2211.07636*, 2022. 4
- [9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 4
- [10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 1, 3, 4
- [11] Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877, 2023. 7
- [12] You Jin Kim, Hee-Soo Heo, Soyeon Choe, Soo-Whan Chung, Yoohwan Kwon, Bong-Jin Lee, Youngki Kwon, and Joon Son Chung. Look who’s talking: Active speaker detection in the wild. *Interspeech*, 2021. 4
- [13] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 1, 3, 4
- [14] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, pages 4953–4963, 2022. 8
- [15] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565, 2019. 3, 4
- [16] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 4, 5
- [17] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhui Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding, 2023. 8, 1
- [18] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. *arXiv preprint arXiv:2303.16058*, 2023. 8
- [19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5
- [20] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4
- [21] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 8
- [22] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 6, 8, 2, 5
- [23] OpenAI. Gpt-4. 2, 4
- [24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 5
- [25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3, 5
- [26] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 8
- [27] Enxin Song, Wenhao Chai, Guan hong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Xun Guo, Tian Ye, Yan

- Lu, Jenq-Neng Hwang, et al. Moviechat: From dense token to sparse memory for long video understanding. *arXiv preprint arXiv:2307.16449*, 2023. 8
- [28] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 2
- [29] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5
- [30] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5
- [31] 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. 8
- [32] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 7, 4, 5
- [33] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *CVPR*, 2016. 1, 3, 4, 8
- [34] 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*, pages 5036–5045, 2022. 2, 5
- [35] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019. 7, 4, 5
- [36] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1, 3, 4
- [37] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, pages 6578–6587, 2019. 4
- [38] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 4, 5

# Shot2Story20K: A New Benchmark for Comprehensive Understanding of Multi-shot Videos

## Supplementary Material

We follow the official template from CVPR for organizing the supplementary material and the indexes of sections, figures and tables are continuous to the main submission for easy reference.

### 5. Annotation Process

In this section, the annotation process for Shot2Story20K is detailed, including single-shot caption annotation in Section 5.1, GPT-4 summary generation prompts in Section 5.2, and human correction of summaries in Section 5.3.

#### 5.1. Human annotation of video-shot captions

Our single-shot video caption annotation process, described in Section 3.2, is a two-phase approach designed for high-quality and style-consistent captions. This procedure also accelerates the annotation process by 3 times. As depicted in Figure 6, the process begins with MiniGPT-4 generating initial captions from structured prompts. While these captions often correctly identify subjects such as “parking lot” and “vehicle”, they sometimes inaccurately describe actions or locations. Annotators then watch the video shot and revise these captions. For instance, errors like “standing in front of”, depicted in red, are corrected to “getting close to the car”, shown in green. Additionally, annotators enrich captions with key details, such as “a close-up shot of the front of a car.” Our single-shot narration caption annotation process follows a similar approach. Differently, we offer ASR text and videos to the annotators and ask them to write down the visually related content and describe the source of the speech. The process has been detailed in Section 3.3.

#### 5.2. GPT-4 summarization prompt

We utilize GPT-4 to summarize video clips, leveraging our detailed video-shot captions and ASR text. The summarization follows a prompt structure adapted from [17], which defines video captions and audio captions for each shot, as depicted in Figure 8. For each video, we organize shot durations, video captions, narration captions, and ASR into a text format (see Figure 7). This arranged content is then fed into GPT-4 for generating the video summary.

#### 5.3. Human correction of video summaries

Our detailed shot captions enable GPT-4 to effectively identify and link subjects across shots, without requiring extra re-identification modules. However, according to human evaluation, about 30% of video summaries struggle to connect objects and scenes across shots. Our annotators review

these summaries alongside the video clips to correct such errors. Figure 9 illustrates this process. While GPT-4 accurately references the same location, such as “the open field” and “the same open field”, it sometimes fails to maintain continuity with elements like “the black car” across scene transitions. Annotators must watch the video and assess the initial summary to make necessary corrections for the final summary. This method ensures the production of high-quality video summaries with efficiency.

### 6. Prompts Used in Our Models

In this section, we elaborate on the prompts used for training and testing our models. We detail the prompts for single-shot video captioning and narration captioning in Section 6.1. For video summarization models like SUM-shot, SUM-holistic, and SUM-text, the prompts are thoroughly explained in Section 6.2.

#### 6.1. Prompt for single-shot captioning

During the training of our single-shot captioning models, we select a random text prompt for each video shot, with different model variants utilizing distinct prompts. The prompts for the single-shot video captioning model that incorporates both visual signals and ASR are depicted in Figure 10, while those relying solely on visual signals are in Figure 11. The prompts used for single-shot narration captioning are displayed in Figure 12. In these figures, bold-faced text, such as “{asr}”, is replaced with specific video information. The arrangement of visual tokens and text prompts, as presented in Figure 3, is not included here for brevity.

#### 6.2. Prompt for video summarization

In Section 3.4, we explore different model variants for video summarization, namely SUM-shot, SUM-holistic, and SUM-text. For training SUM-shot and SUM-holistic, we use the same text prompt as in single-shot video captioning, shown in Figure 10. The key distinction between these models lies in the arrangement of visual tokens: SUM-shot incorporates shot-specific information such as shot number or index along with visual tokens from each shot, whereas SUM-holistic uniformly samples 16 frames across the video. For SUM-text, the input prompt is similar to that of SUM-shot, with the primary difference being the attachment of single-shot captions instead of shot visual tokens. This organized prompt is then fed into Vicuna for both training and testing purposes.

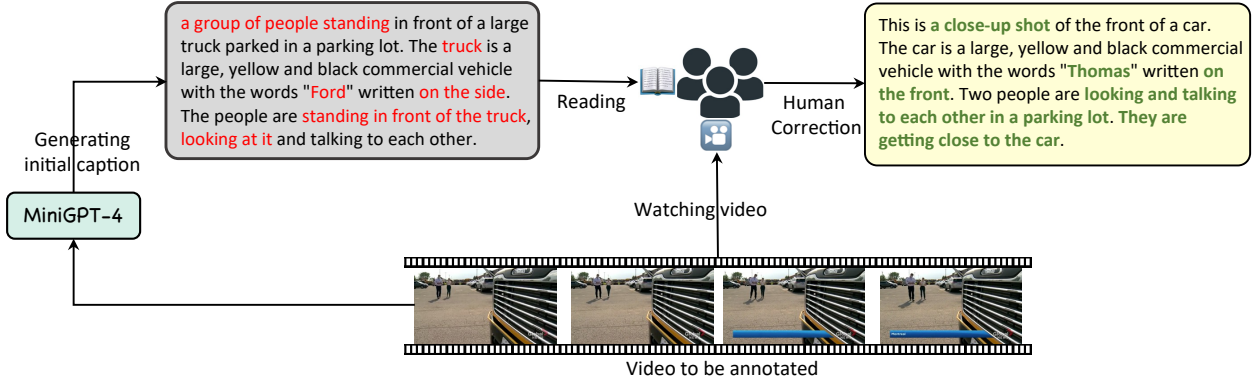


Figure 6. Human annotation process of sing-shot video captioning. Texts in bold green represent correct content, while those in red indicate errors. Please find more explanations in Section 5.1.

The video has **4** shots. It has **15.8 seconds** in total.  
The first action segment starts from **0.0** seconds to **6.5** seconds.  
Visual caption of this clip is: **It's a man in a kitchen chopping something with a knife. The man is wearing a tee shirt with the number 19 on it. He seems to hurt himself and appears in pain. The video is in black and white. The yellow words "DON'T RISK IT" appear on the screen. Then the scene changes to a bowl of guacamole with chips on the side. The video also shows the package.**  
The audio caption of this clip is: **In the audio, the background voice says don't risk getting injured from cutting up avocados.**

The second action segment starts from **6.5** seconds to **10.5** seconds.  
Visual caption of this clip is: **It's a close-up of a person scooping avocado out with a spoon on a wooden cutting board. There is also a white bowl next to it.**  
The audio caption of this clip is: **In the audio, the background voice says the product is refrigerated and pre-scooped for the customers.**

The third action segment starts from **10.5** seconds to **13.5** seconds.  
Visual caption of this clip is: **It's a woman in a pink shirt with a cast on her arm. She holds her thumb up with the injured arm. The word "SWEET" in purple is typed on the screen.**  
The audio caption of this clip is:

The forth action segment starts from **13.5** seconds to **15.8** seconds.  
Visual caption of this clip is: **It's a bowl of guacamole with chips on the side. There is a plastic container of guacamole with a purple package next to it.**  
The audio caption of this clip is:

The ASR of the video is: **ASR content**

Figure 7. Example of textual content for video in Figure 1. Texts in color are specific for input video and are replaced during our generation.

### 6.3. Prompts used for zero-shot QA

In this section, we detail the prompts employed for the zero-shot video question-answering task, which is discussed in Section 3.5.

#### 6.3.1 LLM QA prompt

In our paper, we propose a unique question-answering procedure in which we generate video summaries and prompt an LLM to answer the corresponding question. The text

prompt used for LLM is shown in Figure 13. To better align with the ground truth answers, we prompt the LLM to generate concise answers solely based on the provided video content.

#### 6.3.2 Evaluation prompt

In our paper, we follow the same evaluation procedure as outlined in [22], using ChatGPT-3.5 to assess the alignment of the generated answers with the given ground truth.



You are a chatbot that conducts conversations based on video contexts. You mainly answer based on the given caption contexts from different sources, such as visual caption and audio caption. You can also answer the relevant knowledge of the person or object contained in the video. The video has a high-level topic and the video content is supposed to be coherent. The video can have more than one shot and in each shot different action segments and events exist. Visual caption is provided by a validated annotator and it usually describes more details in the video. Audio caption is the caption of audio in the shot, which is relevant to the visual content and it is provided by a validated annotator. Different shots are annotated separately, but the content and topics of different shots are supposed to be coherent. You should refer to the audio caption and extend the content accordingly. You don't create a video plot out of nothing. The video content is:

**{video content}**

Describe the video in detail and coherently, including the actions and scenes. You should not mention the number of scenes, but you ought to describe the content changes and transition. You ought to describe the content in happening and reasoning order. You should not create a video plot out of nothing.

Figure 8. Prompt template for GPT-4 summarization.

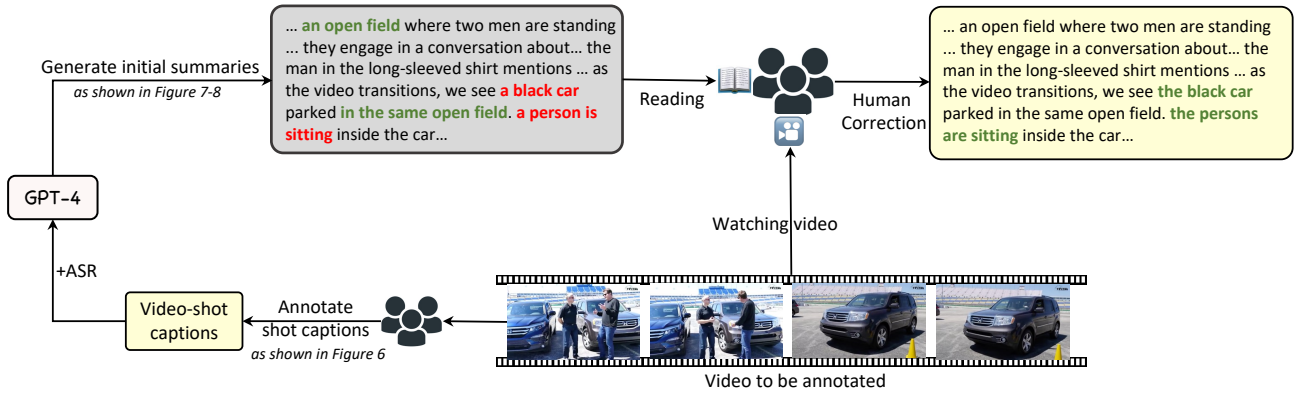


Figure 9. Human correction process of video summaries. Overlapped text is omitted for clarity. Texts in bold green represent correct content, while those in red indicate errors. Please find more explanations in Section 5.3.

## 7. Additional Experiment Results

### 7.1. Comparison of different models

In this subsection, we present an example of video summarization from our Shot2Story20K testing split in Figure 14, using SUM-text, SUM-holistic, and SUM-shot. Both SUM-text and SUM-shot, with their access to shot information including shot count and visual tokens in shots, successfully capture the video’s storyline and transitions. For instance, they accurately depict the sequence involving a woman in the kitchen, almonds in an oven, and the woman speaking to the camera. SUM-text, with its two-stage training, captures more nuanced details found in single-shot captions, like “a bowl of almonds is situated on the counter.” SUM-holistic, while effectively identifying major content and events, falls

short in accurate scene sequencing due to its lack of shot-structured visual tokens. This leads to errors in narrative order, such as reversing the scenes of “shifts back to the woman in the kitchen” and “in the final scene, the woman is seen using a wooden spatula”.

### 7.2. Comparison of different prompts for inference

In this subsection, we evaluate the impact of different prompts on our SUM-shot models for video summarization task, which were utilized during the training process. Table 7 reveals the SUM-shot model’s consistent performance across different prompts, with slight fluctuations ranging from 0.1 to 0.2 in B, M, and R metrics. The most notable variation occurs in C, suggesting a moderate impact of prompt choice.

Random one during training:

1. The audio transcripts are: **{asr}**. Describe this video in detail.
2. In the audio, I hear that: **{asr}**. Take a look at this video clip and describe what you notice.
3. Based on the audio, the speech content is: **{asr}**. Please provide a detailed description of the video.
4. From the audio, I gather the content that: **{asr}**. Could you describe the contents of this video for me?

Figure 10. Prompt during training for single-shot video captioning.

Random one during training:

1. Describe this video in detail.
2. Take a look at this video clip and describe what you notice.
3. Please provide a detailed description of the video.
4. Could you describe the contents of this video for me?

Figure 11. Prompt during training for single-shot video caption without using ASR text.

Prompt	B	M	R	C
Prompt-1	11.7	19.7	26.8	8.6
Prompt-2	11.8	19.7	27.0	7.5
Prompt-3	11.8	19.7	26.9	7.7
Prompt-4	11.7	19.6	26.9	7.8

Table 7. Performance of SUM-shot model with different inference prompts. Prompt IDs correspond to those outlined in Figure 10.

### 7.3. QA summary

In this subsection, we present the results of zero-shot video question-answering using Vicuna, based on textual summaries of video samples from MSRVTT-QA [32] and ActivityNet-QA [35]. Despite the limitations of our summarization model, which scores 8.6 in CIDEr on the Shot2Story20K test split (see Table 4), and the inherent challenges of the videos due to out-of-domain topics or extended durations, the summaries generated from our trained SUM-shot model largely succeed in capturing the key elements of the videos and providing relevant information.

**MSRVTT-QA:** For instance, in Figure 15, `video7089` from MSRVTT-QA portrays a TV show outside the scope of our Shot2Story20K. This genre typically features minimal movement within individual shots, and frequent scene transitions, but a restricted variety of scenes. Yet, our generated summary aptly identifies principal elements such as the judges and contestants, actions like “engaged in a conversation” or “picking up a guitar”, and the setting of an American Idol audition. These details equip the summary to competently address questions from MSRVTT-QA. However, some gaps in detail lead to inaccuracies: of the first 10 questions for `video7089`, 5 are incorrectly answered due to missing information (e.g., Q1, Q3), incorrect summary

content (e.g., Q4), or misalignment with the ground truth (e.g., Q6, Q10).

**ActivityNet-QA:** In Figure 16, we present the video `v_mZYWfmsYQPA` from ActivityNet-QA. The video’s duration is 104 seconds, which is considerably longer than the average in our Shot2Story20K. Our summary effectively identifies important elements such as the main subject’s clothing described as “dressed in a black shirt”, the actions including “speaks to the camera” and “playing table tennis”, and the setting, noted as “in a gym”. These comprehensive details enable the Vicuna model to correctly answer most questions from ActivityNet-QA. Despite these insights, some omissions and discrepancies in the summary contribute to inaccuracies in answering. Specifically, three out of ten questions are answered incorrectly due to either a lack of specific details or mismatches with the ground truth, as seen in questions Q1, Q5, and Q7.

## 8. Additional Implementation Details

**Video-Shot Captioning.** For each video shot, we uniformly sample 4 frames. For testing, a consistent text prompt is used, as depicted in Figure 3. The maximum number of new tokens generated by the LLM is capped at 150 for both training and inference.

**SUM-shot and SUM-holistic.** In SUM-shot, 4 frames per video shot are sampled uniformly; in SUM-holistic, 16 frames per video clip are sampled. The rationale behind sampling 16 frames in a holistic approach is based on our dataset’s average of 4 shots per video, aligning with the SUM-shot approach of 4 frames per shot. For both training and inference, the LLM’s maximum new token count is set at 600. A consistent text prompt, as shown in Figure 4,

Random one during training:

1. The audio transcripts are: {asr}. Describe the audio content of this video in detail.
2. In the audio, I hear that: {asr}. Take a look at this video clip and describe what you notice in the audio.
3. Based on the audio, the speech contents are: {asr}. Please provide a detailed description of the audio in the video.
4. From the audio, I gather the content that: {asr}. Could you describe the contents of the audio in this video for me?

Figure 12. Prompt during training for single-shot narration captioning.

I need your help to identify a specific object, place, person and way in the video based on its description. You must answer my question concisely. The answer is definitely contained in the provided video description. Video content is:

{video summary}

You should answer the question concisely. Based on the video, please answer {video question}

Figure 13. Prompt used for video question answering with summaries.

is used during inference.

For video-shot captioning and video summarization tasks, both models are trained on  $8 \times 2$  A100-80G GPUs using Pytorch. The captioning model is trained for 40 epochs, with the best-performing checkpoint on the validation set used for test performance reporting. To prevent overfitting, text prompts are randomly sampled for each sample, as detailed in Section 6.

**SUM-text.** For tuning the Vicuna 7B model, we arrange our text prompt (see Section 6.2) as a question and the video summary as an answer. During training, we use ground truth single-shot captions. During inference, we use predictions of our trained single-shot video captioning model. The model is trained on 4 A100-80G GPUs over 3 epochs with a learning rate of  $2e-6$ , maintaining other hyperparameters at their default values from the original configuration.

**Video-ChatGPT.** Consistent with SUM-holistic, we uniformly sample 16 frames for both training and inference. Our prompt setup excludes ASR for video summarization, using the video summary as the answer to the question. The training is conducted over 3 epochs with a learning rate of  $2e-6$ , and we retain other hyperparameters at their default settings as specified in the original repository.

**Video Question-Answering with Summary.** For the MSRVT-QA [32] and ActivityNet-QA [35] datasets, we generate video summaries using our trained SUM-shot model, employing only visual tokens during inference. Upon generating these summaries, we integrate them with individual questions from their corresponding videos into the prompt format displayed in Figure 13. This integrated

content is then processed through Vicuna [4] to obtain answers. The evaluation of these results is carried out following the methodology outlined in [22].

## 9. Broader Impact

**Data Limitations and Ethical Considerations.** Our approach does not involve storing or releasing videos; users must download these from original sources and comply with their licenses. Given HD-VILA-100M [34]’s long-standing public availability, we assess a low risk of the currently available videos being removed in the near future. Additionally, our meticulous manual annotation process is designed to avoid any ethical or legal violations.

**Human Rights in Annotation Process.** We have conscientiously structured the annotation process to ensure fair workloads and equitable compensation for annotators, upholding human rights standards.

**Scope of Conclusions.** It is important to recognize that experiments and data, including ours, might only represent a subset of universal realities. Nevertheless, given the wide range of categories covered in our videos, we believe our conclusions offer a robust understanding applicable to various multi-shot video scenarios and durations. These findings, while specific to our dataset, provide significant insight into the broader field of video analysis.

**Usage of Language Models.** Our use of the LLaMA model [29] from Meta is authorized for research purposes. Those intending to use our model post-release should ensure that they have the necessary permissions and adhere to usage restrictions. We express deep respect for the work of developers and contributors, recognizing their integral role



**Human speech:** However, there are a couple of tricks that you need to know in order to be successful. The first is that you must roast your almonds that helps to bring out not only the flavors but the oil as well.

**SUM-text** video summarization:

the video begins with a woman in a green t-shirt standing in a kitchen, holding a blender and engaging in a conversation with the camera. a bowl of almonds is situated on the counter in front of her, indicating that she is likely discussing a recipe that involves almonds. as the video progresses, the scene transitions to a person using a spatula to mix the almonds in a baking tray that is placed in an oven. this suggests that the woman is demonstrating a cooking process, possibly preparing a dish that involves roasted almonds. In the final part of the video, the woman, now wearing a blue top, is seen holding a bowl of almonds and continuing her conversation with the camera. a blender is visible in front of her, suggesting that she is discussing a recipe that involves blending the almonds. the video seems to be a cooking tutorial, with the woman guiding the viewers through the preparation and use of almonds in various dishes.

**SUM-holistic** video summarization:

the video begins with a woman in a well-equipped kitchen, addressing the camera with a smile. she is wearing a green t-shirt and is surrounded by brown wooden cabinets and appliances. as she speaks, she reveals that there are a few tricks that people need to know in order to be successful. the scene transitions to a close-up of a clear, transparent blender. the woman's voice instructs viewers to roast their almonds to bring out not only the flavors but the oil as well. the video then shifts back to the woman in the kitchen, now holding the transparent blender. she continues to speak to the camera, emphasizing the importance of roasting the almonds. in the final scene, the woman is seen using a wooden spatula to scoop up the roasted almonds from a baking tray. she then places the tray in the oven to continue roasting the nuts. throughout the video, the woman's words and actions provide a comprehensive guide on how to successfully roast almonds.

**SUM-shot** video summarization:

the video begins with a woman in a kitchen setting, wearing a blue top and blonde hair. she is holding a clear plastic container in her hands, which contains almonds. as she speaks to the camera, she mentions a couple of tricks that people need to know in order to be successful. the scene transitions to a close-up of a plate of almonds being cooked in an oven using a shovel. the woman's voice can be heard in the background, explaining that roasting the almonds helps to bring out not only the flavors but the oil as well. in the final scene, the woman continues to hold the clear plastic container filled with almonds. she is seated in a well-equipped kitchen, with a blender placed next to her. as she speaks to the camera, she holds a bowl of almonds in her other hand. the video seems to be a tutorial or a cooking demonstration, with the woman sharing her knowledge and tips on roasting almonds.

**Ground truth** video summary:

the video begins with a woman, dressed in green short sleeves, standing in front of a kitchen console. she is engaged in a conversation, presumably about cooking tips, as she stands next to a bowl of almonds and a transparent mixer. as she speaks, she picks up the bowl of almonds, emphasizing the importance of knowing certain tricks for successful cooking. the scene transitions smoothly to a close-up of an oven pan filled with almonds. a wooden spoon is used to stir the almonds around in the pan. the woman's voice continues in the background, explaining the benefits of roasting almonds. she mentions that this process helps to enhance the flavors of the almonds, suggesting that this is one of the tricks she was referring to earlier. the video concludes with the woman once again standing in front of the kitchen console, continuing her discussion. she holds the bowl of almonds in her hands, possibly preparing to demonstrate the next step in her cooking process. throughout the video, the woman's actions and dialogue suggest that she is sharing her knowledge and tips about cooking, specifically focusing on the use of almonds.

Figure 14. Example for video summarization using SUM-text, SUM-holistic and SUM-shot. All three models effectively grasp the video’s main topic and content. However, SUM-shot, with its access to shot information in the input prompt, excels in structuring the storyline compared to SUM-holistic. SUM-text, developed through a two-stage training approach, captures finer details than SUM-shot, highlighting the strengths of the SUM-shot model while also indicating challenges in its training process. Please find more explanation in Section 7.1.

in advancing language modelling and multi-modal learning.

**Future Research and Development.** Aligned with our commitment to the research community and in adherence to CVPR guidelines, we will release both our code and dataset. This is intended to encourage further research and enable others to build upon our work. Although our current experiments require up to  $8 \times 2$  A100-80G GPUs, we are aware this may be a limitation. Consequently, we plan to focus fu-

ture efforts on adapting these experiments to be compatible with a single node of 8 A100 GPUs. It’s important to note that fitting the experiments within an 8 GPU framework is not the primary focus of this paper, but we consider it a crucial step towards making our research more accessible and inclusive for a wider array of research groups.





**SUM-shot** video summarization:

the video begins with a man in a black shirt and a woman in a blue shirt sitting behind a blue and silver-colored table with a city skyline in the background. they are engaged in a conversation, with the man occasionally clapping his hands. the scene then transitions to two men standing next to a large screen with american idol written on it. one man, dressed in a white shirt and black pants, picks up a guitar and starts playing. the other man, wearing a black shirt and khaki pants, claps his hands and smiles. in the final scene, the man in the white shirt and black pants continues to play the guitar, while the woman in the blue and silver-colored shirt looks at the camera. the man in the black shirt and khaki pants claps his hands and smiles. the video captures the interactions and performances of these individuals in front of the american idol backdrop.

**Questions:**

- Q1: who speaks to tv show judges?
- Q2: who do three judges talk to?
- Q3: what is a black male laughing on?
- Q4: what talks to a performer?
- Q5: what did the man audition for?
- Q6: who does a panel talk to?
- Q7: what do judges talk on?
- Q8: what shows judges of a singing competition?
- Q9: how many contestants are standing before the judges at american idol?
- Q10: what is going on?

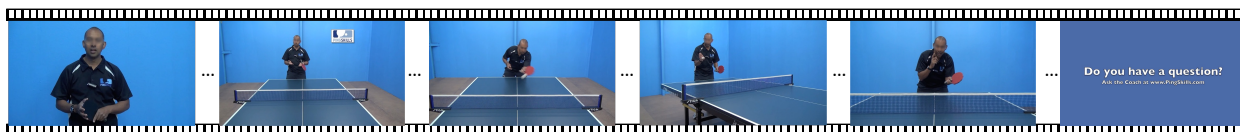
**Answers:**

- GT1: person
- GT2: contestant
- GT3: idol
- GT4: panel
- GT5: idol
- GT6: performer
- GT7: idol
- GT8: video
- GT9: two
- GT10: reality

**Vicuna answers using summaries & GPT evaluation:**

- A1: Based on the video description, it is not clear who is speaking to TV show judges. ❌
- A2: The three judges in the video are the man in the black shirt and khaki pants, who claps his hands and smiles, and the woman in the blue and silver-colored shirt, who looks at the camera. ✅
- A3: I'm sorry, but I cannot identify a specific person laughing in the video, as there is no mention of a person laughing in the description. The video appears to be focused on the interactions and performances of the individuals in front of the American Idol backdrop. ❌
- A4: The man in the white shirt and black pants is a performer. ❌
- A5: The man in the white shirt and black pants auditioned for American Idol. ✅
- A6: Based on the video, it appears that a panel talk is happening between the man in the black shirt and the woman in the blue shirt, who are sitting behind a table with a city skyline in the background. The man occasionally claps his hands, and the scene transitions to two men standing next to a large screen with "American Idol" written on it. One man, dressed in a white shirt and black pants, picks up a guitar and starts playing, while the other man, wearing a black shirt and khaki pants, claps his hands and smiles. In the final scene, the man in the white shirt and black pants continues to play the guitar, while the woman in the blue and silver-colored shirt looks at the camera. The man in the black shirt and khaki pants claps his hands and smiles. ❌
- A7: Judges talk about performances in front of the American Idol backdrop. ✅
- A8: The video shows judges of a singing competition. ✅
- A9: Two contestants are standing before the judges at American Idol. ✅
- A10: It appears that the video is a behind-the-scenes look at a recording session for a musical performance, possibly for a talent competition such as American Idol. The individuals shown in the video are a group of musicians who are preparing for a performance in front of a panel of judges. The man in the black shirt and khaki pants is likely a music producer or judge, while the man in the white shirt and black pants is likely a contestant on the show. The woman in the blue and silver-colored shirt is likely another member of the group or a crew member. ❌

Figure 15. Example for video question-answering using summaries - MSRVTT. We show the zero-shot video summary using our trained SUM-shot model. The summary can capture most of the video, such as the subject, the action and the transitions. We also show the answers from Vicuna to the first 10 questions and the evaluation results from GPT. Please find more explanations in Section 7.3.



v\_mZYWfmsYQPA.mp4 duration: 104 seconds

#### **SUM-shot** video summarization:

the video features a man dressed in a black shirt, playing table tennis against a blue wall in a gym. as he plays, he speaks to the camera and gestures with his hands. the scene then transitions to a close-up shot of the man, still in his black shirt, playing table tennis against a blue wall in a gym. he continues to speak to the camera and gestures with his hands. following this, the video shifts to a close-up shot of the man playing table tennis against a blue wall in a gym. he continues to speak to the camera and gestures with his hands. in the final scene, the video displays a blue background with white text that reads, do you have a question? ask the coach at [www.fingskill.com](http://www.fingskill.com). the man in the black shirt stands in front of the blue wall, still speaking to the camera and gesturing with his hands.

#### **Questions:**

- Q1: what is the person in the video doing?
- Q2: is the person in the video wearing a watch?
- Q3: is the person in the video wearing a hat?
- Q4: what is in front of the person in the video?
- Q5: what happened before playing table tennis?
- Q6: where is the video taken?
- Q7: how safe is this sport in the video?
- Q8: what benefit can this exercise bring?
- Q9: how many people are there in video?
- Q10: what color of the clothes does the person wear in the video?

#### **Answers:**

- GT1: play table tennis
- GT2: yes
- GT3: no
- GT4: table tennis table
- GT5: introduction of pang ball
- GT6: indoor
- GT7: security
- GT8: reaction capacity
- GT9: 1
- GT10: black

#### **Vicuna answers using summaries & GPT evaluation:**

- A1: The person in the video is playing table tennis against a blue wall in a gym, and speaking to the camera while gesturing with his hands. ✓
- A2: Based on the video, the person in the video is not wearing a watch. ✗
- A3: No, the person in the video is not wearing a hat. ✓
- A4: In the video, there is a man dressed in a black shirt playing table tennis against a blue wall in a gym. ✓
- A5: Based on the video, it is not clear what happened before the man started playing table tennis. ✗
- A6: Based on the video, it appears to be taken in a gym. ✓
- A7: Based on the video, it is difficult to determine the safety of the sport featured. The video only showcases a man playing table tennis against a blue wall in a gym, but does not provide any information on the safety precautions or measures taken to ensure the safety of the players. ✗
- A8: Based on the video, the benefit of this exercise is to improve hand-eye coordination and table tennis skills. ✓
- A9: Based on the video, there is only one person present in the video. ✓
- A10: The person in the video wears a black shirt. ✓

Figure 16. Example for video question-answering using summaries - Activity Net. We show the zero-shot video summary using our trained SUM-shot model. The summary can capture most of the video, such as the subject, the action and the transitions. We also show the answers from Vicuna and the evaluation results from GPT. Please find more explanations in Section 7.3.