SHOT2STORY: A NEW BENCHMARK FOR COMPREHENSIVE UNDERSTANDING OF MULTI-SHOT VIDEOS

Anonymous authors
Paper under double-blind review

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 Shot2Story with detailed shot-level captions, comprehensive video summaries and question-answering pairs. 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 captioning, multi-shot video summarization, and multi-shot video question answering. Preliminary experiments show some challenges to generate a long and comprehensive video summary for multi-shot videos. Nevertheless, the generated imperfect summaries can already achieve competitive performance on 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 contain multiple events, dense captioning benchmarks (Ego4D (Grauman et al., 2022), YouCook2 (Zhou et al., 2018), ActivityNet-Caps (Krishna et al., 2017)) capture information of multiple events in videos ranging from 3-20 minutes. However, even within seconds, more than one event often occurs in daily videos such as news broadcasts, 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 (Cutting et al., 2011). Although some existing captioning benchmarks (Xu et al., 2016; Krishna et al., 2017; Zhou et al., 2018) 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 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 relate 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 Shot2Story for audio-visual understanding of multi-shot videos. We collect a dataset of 42,958 short videos where the average number of shots in each video is 4.4. 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 (OpenAI) 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 example of one annotated video is shown in Figure 1.

To benchmark the advances of multi-modal video understanding, we designed several distinctive tasks using our dataset, including single-shot video captioning, multi-shot video summarization, and multi-shot video question answering. We design and implement several baseline models using a

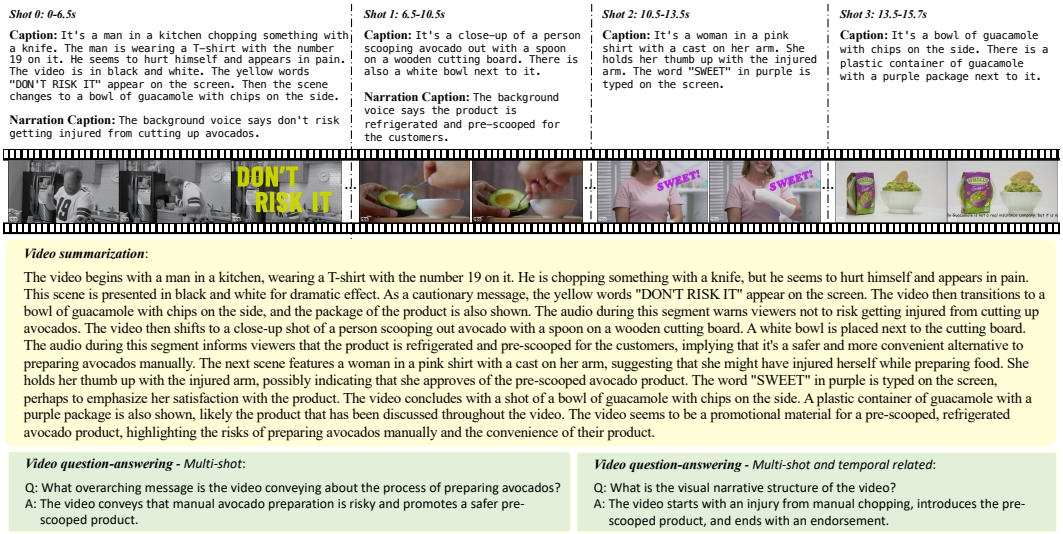

Figure 1: An annotated example of our Shot2Story with sing-shot visual captions and narration captions. Moreover, we provide coherent and reasonable video summaries, and question-answering pairs to facilitate comprehensive understanding of multi-shot videos.

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 joint understanding of visual and audio content, (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 multi-shot summaries can be used on the proposed multi-shot QA benchmark and generalized to other datasets with longer durations (ActivityNet(Krishna et al., 2017)) and out-of-domain topics (MSRVTT(Xu et al., 2016)), validating the quality of our annotated summaries. Without any bells and whistles, we attain competitive results on zero-shot video question-answering by converting the problem into pure text-based QA with the generated video summaries.

2 THE SHOT2STORY BENCHMARK

Our new benchmark Shot2Story contains 42,958 videos. The length of each video is ranging from 10s to 40s. We first use an off-the-shelf shot detection method TransNetV2 (Souček & Lokoč, 2020) to split each video 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.

2.1 DATA PREPARATION

We source videos for our dataset from the public video benchmark HDvila100M (Xue et al., 2022). It offers a large collection of narrative videos, comprising 3M YouTube videos segmented into 100M clips, each about 13 seconds long. We 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 (Radford et al., 2021) encoders. We only keep the videos containing at least one shot that is visually correlated to ASR with

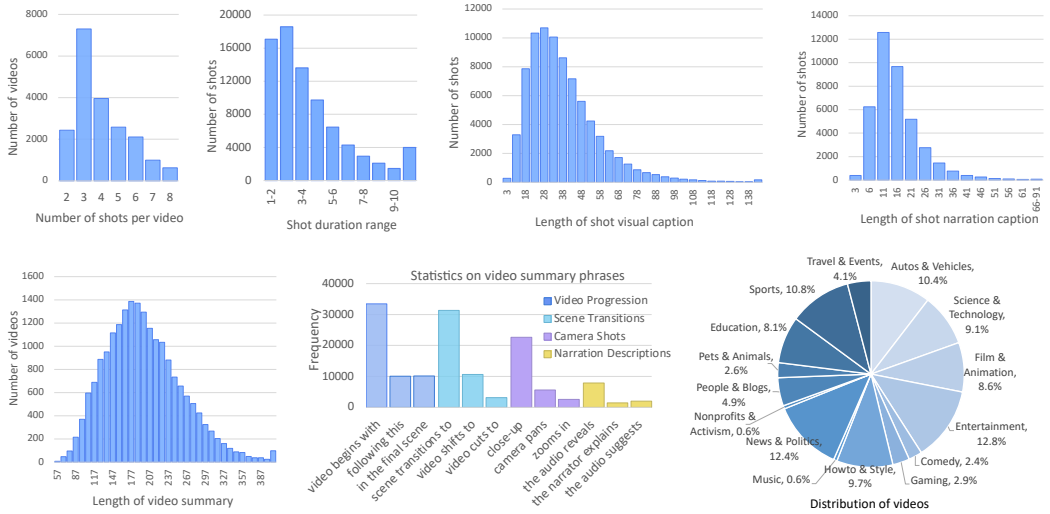

Figure 2: Statistics of Shot2Story . Our dataset features detailed visual captions and narration captions, and video summaries, highlighting video progressions, transitions, camera cuts and narration descriptions, with statistics of frequent expressions depicted in the figure.

a threshold of 0.25. Next, 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 change detector in PySceneDetect[1] with a low threshold of 11 on our segmented shots. If the filter is unable to detect scene changes at this 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 2.1M sampled video clips from HD-VILA-100M, we obtain $42,958$ 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.

2.2 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 these two caption annotation to facilitate separate modeling of these two types of information source. For visual-only captions, 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 (Zhu et al., 2023) by sampling 4 image frames from the video clip and prompting the model using below 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 short video clips, both comprehensively and reasonably. It is able to describe different subjects including persons, animals, food, tools, and virtual objects like animated characters. Annotators are first instructed to correct any errors in the original captions. The mistakes include incorrect descriptions of the object categories, attributes, actions, facial expressions, etc. Also, there might be some subjective descriptions generated by MiniGPT-4 such as emotion and atmosphere. We ask annotators to remove all these subjective descriptions. We then ask annotators to supplement the information about the major subjects, actions, and backgrounds present in the video. The goal is for the resulting captions to accurately capture the key elements of each video shot. Statistics shows over 80% of single-shot visual captions are manually corrected. For narration captions, annotators should

[1]https://www.scenedetect.com/

Table 1: High-level comparison of our dataset to previous ones. The summary length of ActivityNet and YouCook2 are combined length of captions in one video. M and G denote manual and generated.

Dataset	Annotation	Multi-shot Video	Multi-event Descriptions	Detailed Summary	Summary Length	#Videos	Avg. Duration
MSRVTT (Xu et al., 2016)	M	✓	✗	✗	-	10K	15s
ActivityNet Caps (Krishna et al., 2017)	M	✓	✓	✗	52.4	20K	3min
VideoStorytelling (Li et al., 2019)	M	✓	✓	✓	162.6	105	12.5min
Ego4D (Grauman et al., 2022)	M	✗	✓	✗	-	10K	23min
YouCook2 (Zhou et al., 2018)	M	✓	✓	✗	67.8	2K	6min
VAST (Chen et al., 2024)	G	✓	✗	✓	32.4	27M	5∼30s
Shot2Story	M+G	✓	✓	✓	218.3	43K	17.1s

watch the video and interpret the audio content that visually correlates into descriptive narration, including information sources. For example: "According to the woman in white, the room is not very clean." All narration captions are manually drafted from scratch. An example of this annotation process is shown in Appendix A.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 coherent style and tend to cover more details of the video.

In contrast to the traditional video captioning benchmarks (Xu et al., 2016; Krishna et al., 2017; Zhou et al., 2018), we also annotate narration captions in addition to the visual-only captions. Different from existing audio captioning benchmarks (Gemmeke et al., 2017), we focus more on human speeches rather than acoustic events. 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 annotators need 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 (Kim et al., 2021) and visual grounding (Anne Hendricks et al., 2017; Zhou et al., 2019), but they only work well on constraint scenarios.

2.3 Annotation of Video Summary

To create video summaries with 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 (OpenAI) to generate a cohesive summary. The text prompt we use is shown in Appendix A.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" and "returns to the scene" to denote recurring scenes across 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 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 summaries. Since our shot-level captions for generating video summaries are manually checked and annotated, the initial video summaries merely have factual errors, with exceptions for some subject identity and scene mismatches. We require annotators to pay more attention to these errors and ensure holistic summary is accurate and comprehensive. Statistics show that over 40% summaries are manually corrected.

Despite the rigorous verification process, the reliance on automated generation introduces certain inaccuracies and biases. Common pitfalls in the generated summaries include the omission of minor yet contextually important details and a bias towards emphasizing more prominent actions or objects, potentially overlooking less conspicuous elements. This is partly due to our dataset predominantly featuring human-centric activities, as a result of our video filtering process that selects videos rich in visually related audio information and sourcing from HDVILA Xue et al. (2022), which primarily curates content related to human activities. Consequently, our annotations tend to highlight salient

events and large-scale objects essential to the video's storyline, mentioning smaller objects only when they directly contribute to the narrative.

2.4 ANNOTATION OF QUESTION-ANSWERING PAIRS

We annotate the question-answering pairs on videos in validation and testing splits. To construct this benchmark, we begin with the human-annotated video summaries from Section 2.3, which is detailed in video content. We then prompt GPT-4 (Achiam et al., 2023) to generate candidate question-answer pairs in three predefined categories: temporal-related (*e.g.*, *Does the woman appearing at the end of the video wear any accessories?*), multi-shot holistic understanding (*e.g.*, *What is the overarching theme of the video?*), and audio content related (*e.g.*, *Which specific car model does the background voice mention, and what visual features confirm its identity?*). The text prompt we use is shown in Appendix B.3.

The quality is then assessed through further automatic filtering and manual checks. Annotators are instructed to verify each question-answer pair against the video content and discard any with mistakes. Simultaneously, they are asked to categorize the questions, where a single question might fall into multiple categories, facilitating evaluating different aspects of multishot understanding *i.e.*, understanding sequences of events and actions (temporal related), integrating information across multiple shots (holistic understanding), correlating audio content with visual elements (audio related), and others (directly discarded). After this verification, annotators carefully review the quality of the questions to ensure they solely correlate with the video content and that the answers are not revealed in the question texts. This step includes a thorough manual check to address potential mistakes and biases from the initial verification. Through these sequential annotation stages, we ensure high-quality annotations. The process starts with detailed manual single-shot captioning, followed by careful verification and correction at each stage. Even with LLMs introduced to reduce workload, the human-involved process and thorough procedures keep the resulting annotations well-aligned with human labeling.

Subsequently, to optimize clarity and reduce unnecessary information in the QA pairs, we remove QA pairs with questions exceeding 28 words or answers exceeding 20 words. Then, we employ Vicuna-13B (Chiang et al., 2023) to attempt answering the questions without video context and discard those answerable without accessing the videos. Finally, the pairs are tested against popular methods such as Video-LLaVA (Lin et al., 2023), LLaMA-VID (Li et al., 2023e), Video-ChatGPT (Maaz et al., 2023), and VideoChat2 (Li et al., 2023c). Questions correctly answered by more than two models are excluded to guarantee that our dataset poses a substantial challenge. Finally, we obtain 4905 QA pairs for validation and 6465 QA pairs for testing set.

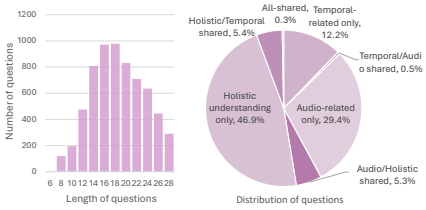

Figure 3: Distribution of multi-shot video QA benchmark. Questions from different categories overlap. *All-shared* means questions fall under all three categories.

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 MSRVTT (Xu et al., 2016), YouCook2 (Zhou et al., 2018) and ActivityNet Caps (Krishna et al., 2017), also use multi-shot videos as annotation source, but they either annotate a holistic caption for the video (MSRVTT) 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, motivating us to annotate distinct captions for each shot. Ego4D (Grauman et al., 2022) only annotates dense visual captions but not audio captions for relatively long egocentric videos. Video Storytelling (Li et al., 2019) 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 (Chen et al., 2024) feeds generated video and audio captions into an LLM to generate video summary. However, it processes multi-shot video as a whole and lacks the granularity

Table 2: Performance of models on video shot captioning using different modalities, following the settings of VAST(Chen et al., 2024). The models are fine-tuned on Shot2Story video shot captions. V, A and S are abbreviated for vision, audio and subtitle (ASR text) respectively.

Model	FT Modality	B4	M	R	C
VAST	V+S	10.7	16.1	30.3	33.8
VAST	V+A+S	10.7	16.1	30.4	34.0
MiniGPT4-C	V	9.2	14.7	27.9	25.1
MiniGPT4-C	V+S	11.8	16.7	30.1	35.9
VideoChat2-C	V	8.8	16.1	27.9	23.7
VideoChat2-C	V+S	**13.8**	**18.7**	**32.1**	**43.9**

Table 3: Performance of models on video summarization. The models are fine-tuned on Shot2Story video summaries. V and S are abbreviated for vision and subtitle (ASR text) respectively.

Model	FT Modality	B4	M	R	C
Video-ChatGPT w/o ASR (Maaz et al., 2023)	V	4.8	17.3	21.3	1.5
Video-ChatGPT (Maaz et al., 2023)	V+S	3.6	17.8	19.7	1.0
MiniGPT4-SUM-holistic	V+S	7.8	16.9	23.4	2.8
MiniGPT4-SUM-shot w/o ASR	V	10.4	18.5	25.8	4.8
MiniGPT4-SUM-shot	V+S	12.4	19.7	27.6	7.6
VideoChat2-SUM-shot	V+S	**12.7**	**19.8**	**28.3**	**9.0**

of the events in different shots. Moreover, VAST directly uses predicted captions without any human verification, leading to potentially noisy and biased summaries towards the captioning models. Our dataset stands out from VAST with its accurately annotated visual and narration shot captions. 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 218.3 words for the video summary, which is much longer than existing benchmarks, and is longer than the combined length of captions in one video in ActivityNet and YouCook2.

Furthermore, our Shot2Story-QA introduces unique and complex challenges through its emphasis on shot transitions and multi-event progression, setting it apart from benchmarks like MSRVTT-QA Xu et al. (2017) and ActivityNet-QA Yu et al. (2019). For instance, unlike existing benchmarks Xu et al. (2017); Yu et al. (2019) that assess understanding at a single time point, *e.g.*, "Who do three judges talk to?" (MSRVTT-QA, in Figure 19), or general inquiries like "What is the person in the video doing?" (ActivityNet-QA, in Figure 20), Shot2Story-QA includes "Temporal-related" and "Multi-shot Holistic Understanding" questions. A temporal-related question, such as "What is the man's immediate action after handling the skewer?" shown in Figure 21, requires models to comprehend the sequence and progression of events, linking consecutive actions meaningfully. Similarly, multi-event progression questions like "How does the setting change from the start to the end of the video?" necessitate understanding multiple concurrent events within their temporal context. This provides a more rigorous and nuanced evaluation framework for temporal understanding.

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 36951, 1982 and 4025, respectively. We resize the frames to 224×224. We adapt two popular VLMs to accomodate our tasks: MiniGPT-4 (Zhu et al., 2023) and VideoChat2 (Li et al., 2023c). For MiniGPT-4, we employ ViT-G/14 (Fang et al., 2022) and Q-Former (Li et al., 2023a) as visual encoder, and Vicuna v0-7B (Chiang et al., 2023) as the language model. We load pretrained Q-Former and MLP from MiniGPT-4 (Zhu et al., 2023). In training, we update only Q-Former and MLP parameters, keeping the ViT and LLM frozen. For VideoChat2, we employ UMT-L(Li et al., 2023d) as backbone and load pretrained Q-Former and MLP from VideoChat2 (Li et al., 2023c). During training, we adopt LoRA(Hu et al., 2021) and AdamW (Loshchilov & Hutter, 2017) with a learning rate of 8e-5. We train both models for 10 epochs with a batch size of 128 for single-shot video captioning. We finetune our video summarization models

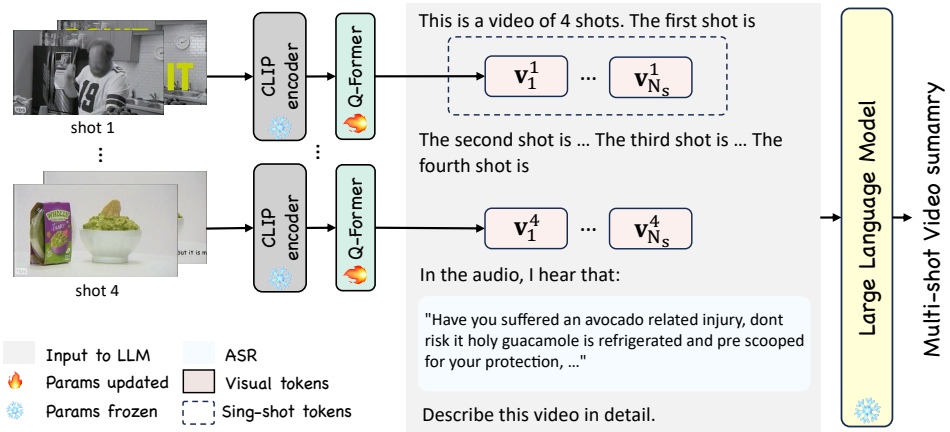

Figure 4: Model structure for multi-shot video summarization model SUM-shot. We arrange visual tokens sequentially for each single shot and in a multi-shot format to encapsulate multi-shot information. Additionally, ASR text is incorporated for audio-visual video summarization.

on the single-shot captioning models with a batch size of 32. For captioning and summarization, we evaluate the models using BLEU@4 (Papineni et al., 2002) (B), METEOR (Denkowski & Lavie, 2014) (M), ROUGE (Lin, 2004) (R), and CIDEr (Vedantam et al., 2015) (C).

3.2 Single-shot video captioning

This task involves generating descriptions for individual video shots, where the target description is a concatenation of the visual-only and narration caption for a video shot. This task requires a joint understanding of visual and speech information. Specifically, we adopt a similar structure as we generate pseudo captions for data annotation in Section 2.2. First, we sample N_s frames from a video shot, encode them using a fixed vision encoder, then feed the encoded features to a Q-Former to produce visual tokens. Further, we combine the visual tokens and an optional ASR text into a unified LLM prompt to facilitate both visual and narration understanding. We adapt the framework of MiniGPT-4 (Zhu et al., 2023) and VideoChat2 (Li et al., 2023c), with the two models denoted as MiniGPT4-C and VideoChat2-C for brevity. We compare with baseline model VAST (Chen et al., 2024), which is able to processes audio, vision, and subtitle inputs simultaneously.

The results are shown in Table 2. Benefiting from extensive pretraining, VAST achieves 34 on C and 30 on R, comparable to MiniGPT4-C. Incorporating an additional audio modality results in a negligible performance difference, indicating that audio content only contributes marginally given the ASR text as input on our benchmark. We then assess variants of our models, MiniGPT4-C and VideoChat2-C, with and without the additional ASR text. It shows that including ASR texts significantly enhances performance across all metrics, with a notable boost in R and C, highlighting the relevance of audio content to our video captions. Furthermore, VideoChat2-C, featuring a superior visual backbone and extensive video pretraining, consistently outperforms MiniGPT4-C. This superiority highlights the importance of advanced visual backbone and video pretraining, confirming our benchmark's robustness. However, despite these advances, the results also indicate room for improvement. Figure 5 (a) showcases output examples of our model's single-shot video captioning, detailing visual elements and audio content effectively, capturing actions like "gesturing to explain her fear" and secondary elements such as "a red stuffed doll next to her".

3.3 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. Due to the complexity of this task, we adopt GPT-V (Achiam et al., 2023) to generate a supplementary training set, with video summaries for another 90K videos, sampled in the same approach as described in Section 2.1. Please check the annotation prompt and data samples in Appendix B.5. First, we finetune an existing video caption model Video-ChatGPT (Maaz et al., 2023) by instruction-tuning it on our video summary data, with and without the additional ASR text input. Then, we experiment with three different architecture designs based on MiniGPT4. The first model MiniGPT4-SUM-holistic uses a similar

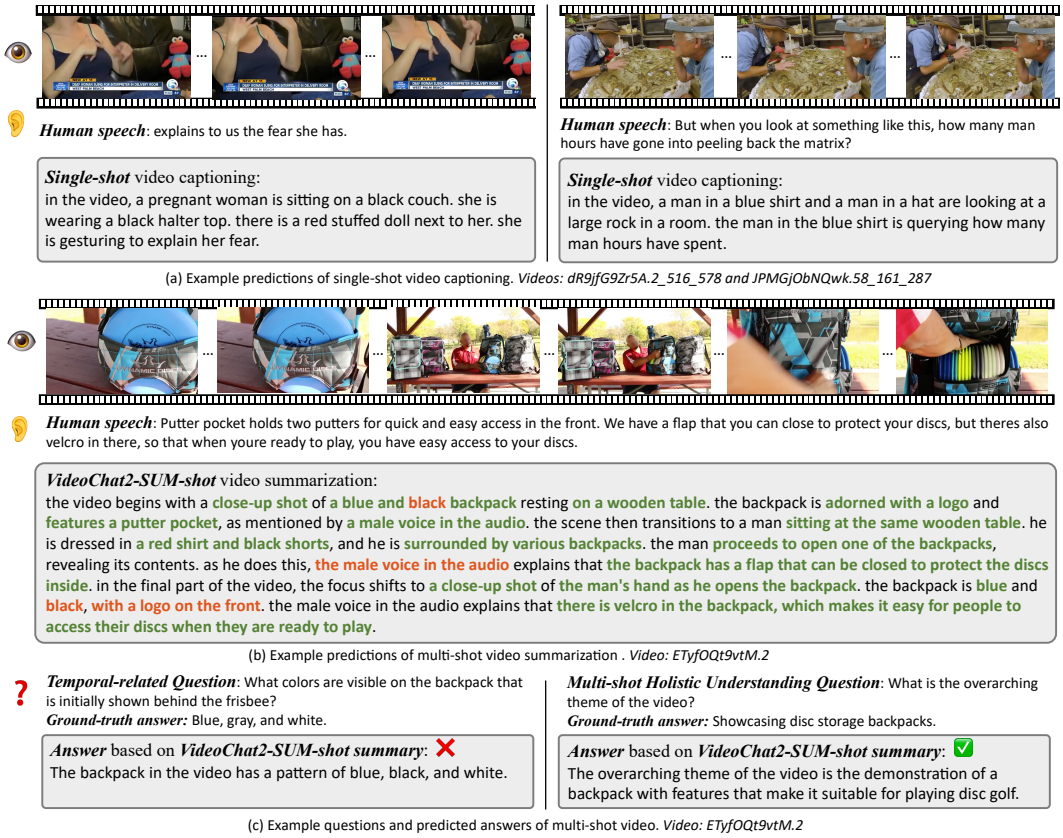

Figure 5: Example predictions of our models. (a) shows single-shot video captioning results of VideoChat2-C, which incorporates audio and visual content correctly (b) shows multi-shot video summarization of VideoChat2-SUM-shot, with accurate descriptions in green and errors in red, illustrating the model's ability to narrate event sequences (c) shows two sample questions of the video in (b). The answers are based on the generated summary of VideoChat2-SUM-shot.

pipeline as MiniGPT4-C. We uniformly sample 16 frames from the full video clip and prompt the LLM with frame tokens and ASR text. The second model MiniGPT4-SUM-shot w/o ASR, neglecting ASR input, 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. The third model, MiniGPT4-SUM-shot further incorporates ASR text input as an additional input. Further, we replace the backbone of MiniGPT4-SUM-shot with the more advanced VideoChat2 model, resulting in the VideoChat2-SUM-shot model variant. Compared to SUM-shot, SUM-holistic does not have explicit shot information and relies on the LLM to parse the video shots using the provided frame features.

Table 3 shows the results of the models. It is shown that MiniGPT4-SUM-holistic is worse than MiniGPT4-SUM-shot, showing the importance of the shot structure in predicting a video summary matching the transition of shots. MiniGPT4-SUM-shot w/o ASR underperforms compared to MiniGPT4-SUM-shot and outperforms MiniGPT4-SUM-holistic, highlighting the significance of both audio information and shot structure in multi-shot video understanding. Compared VideoChat2-SUM-shot and MiniGPT4-SUM-shot, the former model achieves the best performance, indicating the benefit of advanced vision backbone and video pretraining. Video-ChatGPT obtains much worse performance comparing to our models, potentially due to their weakness in processing multiple scenes and building the correlation between visual frames and ASR texts. It directly encodes the whole video into a sequence of tokens, potentially losing significant frame details and essential correlation between ASR and visual frames, while ours directly feed frames tokens into the LLM without compressing them.

Figure 5 (b) showcases predictive capabilities of our VideoChat2-SUM-shot model. The model adeptly narrates event sequences with appropriate emphasis. For instance, it details the backpack's colour and location, and rationalizes the item in the beginning shot, *i.e.*, "putter pocket", aligning

VideoChat2-SUM-shot video summarization:

in the opening scene of the video, a man stands on a ping pong table in an indoor sports facility. he is dressed in a black polo shirt with a logo on the left chest, and he holds a ping pong racket in his right hand. the background is a vibrant blue, and there's a sign that reads welcome to the world of ping pong. as he speaks, he introduces the topic of the video, which is about the basics of ping pong. transitioning to the second scene, the same man continues to address the audience, now holding a red ping pong ball in his hand. he emphasizes the importance of understanding the basics of ping pong, suggesting that this is a fundamental aspect of the sport. in the third scene, the man is seen again on the ping pong table, this time preparing for a serve. he demonstrates the proper grip and stance for a serve, which is a crucial part of the game. the video then cuts to the fourth scene, where a different man appears on the ping pong court. he is wearing a white t-shirt emblazoned with the website www.pingponginfo.com and the text ping pong information . this man speaks about the importance of proper footwork in ping pong, highlighting the role it plays in the game. finally, in the fifth and last scene, the same man from the first scene reappears. he is still on the ping pong table, and he poses a question to the viewers, asking if they have a question about ping pong. this serves as an invitation for the audience to engage with the content and potentially learn more about the sport. throughout the video, the speakers, including the man on the ping pong table and the man in the white t-shirt, provide informative commentary on the basics of ping pong, emphasizing the importance of understanding the fundamentals of the game.

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?

Answers:

GT1: play table tennis
GT2: yes
GT3: no
GT4: table tennis table
GT5: introduction of pang ball

Vicuna answers using summaries & GPT evaluation:

A1: The person in the video is demonstrating the basics of ping pong, including proper grip, stance, serve, and footwork. ✅
A2: No, 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, the person is standing on a ping pong table in an indoor sports facility. ✅
A5: Based on the video, it is not clear what happened before the man started playing table tennis. ❌

ActivityNet-QA case: v_mZYWfmsYQPA.mp4 *duration: 104 seconds*

Figure 6: Example predictions of our model on zero-shot question answering an ActivityNet-QA video. More questions and explanations can be found in Appendix C.2.

with the ASR with "by a male voice in the audio". However, some predictions that marked in red are erroneous, such as the incorrect "black" colour and the non-existent "with a logo on the front" in the ending shot. These inaccuracies likely stem from the LLM's tendency to "hallucinate" plausible yet non-factual details. Despite these errors, the model demonstrates proficiency in generating consistent and nuanced summaries, highlighting our model's potential and the challenges our dataset presents.

3.4 Video question-answering with video summary

Generated video summaries are supposed to be grounded and detailed, covering rich elements like event progression, holistic topics and audio elements, making them suitable for other vision tasks such as video question-answering. Existing work (Guo et al., 2023; Zhang et al., 2023) uses image or video frame captions as input to an LLM to generate question responses. However, little work has been done for the capacity of video summaries. We directly apply our video summarization model on video QA benchmarks, *i.e.* MSRVTT-QA (Xu et al., 2017), ActivityNet-QA (Yu et al., 2019) and our Shot2Story-QA.

Specifically, we first split the testing videos into video shots, and then feed the videos into our SUM-shot models. The generated summaries and the associated questions are then fed into a Vicuna model to derive the answers with the prompt shown in Appendix B.4.1. 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 responses generated by LLM tend to be long sentences, we leverage the `gpt-3.5-turbo` model to generate a binary decision on whether the answer is correct, following Video-ChatGPT (Maaz et al., 2023).

Zero-shot video question-answering. As shown in Table 4, our results with VideoChat2-SUM-shot surpass 5 out of 6 existing video-VLMs on MSRVTT-QA and 4 out of 6 existing models on ActivityNet-QA. Furthermore, our results are comparable to the SOTA performance on MSRVTT-QA with Video-LLaVA (Lin et al., 2023). Note that these models require extensive instruction-tuning data to learn to directly generate answers from visual features and the text prompt, whereas our model bypasses instruction tuning by distilling the video information into a summary. Our model also follows the zero-shot QA setting since the model only uses Shot2Story as training data. Note that MSRVTT contains a large portion of videos with out-of-domain topics such as TV shows (e.g., Figure 19) and food, while ActivityNet has much longer videos than our training videos (e.g., Figure 6. 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

Table 4: Performance on video question answering on MSRVTT-QA and ActivityNet-QA. IT means whether the model uses video-text instruction tuning data. All methods follow the zero-shot manner.

Model	Training Data	IT	QA Input	MSRVTT QA	ActivityNet QA
VideoChat (Li et al., 2023b)	Cap.+QA	✓	V+T	45.0	26.5
Video-ChatGPT (Maaz et al., 2023)	Cap.+QA	✓	V+T	49.3	35.2
MovieChat (Song et al., 2023)	Cap.+QA	✓	V+T	52.7	45.7
LLaMA-VID (Li et al., 2023e)	Cap.+QA	✓	V+T	57.7	47.4
VideoChat2 (Li et al., 2023c)	Cap.+QA	✓	V+T	54.1	**49.1**
Video-LLaVA (Lin et al., 2023)	Cap.+QA	✓	V+T	**59.2**	45.3
MiniGPT4-SUM-shot	Summary	✗	T	57.7	45.6
VideoChat2-SUM-shot	Summary	✗	T	58.5	47.1

Table 5: Benchmark on Shot2Story-QA. IT means usage of video-text instruction tuning data. Summary, Cap. and QA denote video summary, captions and question-answering pairs.

Model	Training data	IT	QA Input	Temporal related	Holistic understanding	Audio related	Overall
LLaMA-VID (Li et al., 2023e)	Cap.+QA	✓	V+T	7.9	9.7	11.4	9.7
Video-ChatGPT (Maaz et al., 2023)	Cap.+QA	✓	V+T	13.1	15.5	14.3	14.2
VideoChat2 (Li et al., 2023c)	Cap.+QA	✓	V+T	15.1	15.4	13	14.5
Video-LLaVA (Lin et al., 2023)	Cap.+QA	✓	V+T	16.4	14.8	11.7	14.3
MiniGPT4-SUM-shot	Summary	✗	T	28.9	31.9	36.7	32.5
VideoChat2-SUM-shot	Summary	✗	T	36.1	41.5	43.8	40.5

summary is a high-quality abstraction of the video, facilitating a wide range of tasks including video QA and video-based conversation. Moreover, our model achieves strong results on ActivityNet-QA, which predominantly consists of single-shot long videos, validating that models trained with multi-shot videos can effectively generalize to single-shot videos.

Multi-shot video question-answering. As shown in Table 5, we benchmark existing and our proposed video summary models on Shot2Story-QA. Specifically, four popular video-VLMs are compared, *i.e.*, Video-ChatGPT (Maaz et al., 2023), LLaMA-VID (Li et al., 2023e), VideoChat2 (Li et al., 2023c) and Video-LLaVA (Lin et al., 2023). The predicted summaries of MiniGPT4-SUM-shot and VideoChat2-SUM-shot are used to tackle the QA task with the same configuration as zero-shot VQA. Accuracies on temporal-related, multi-shot holistic understanding and audio-related are reported, with the "overall" metric showing the average score from these three sub-tasks. The current video-VLMs present unsatisfying results, potentially due to two factors: (1) The current models does not have audio or ASR as input, lacking capacity with audio-related understanding. (2) Current models do not have training data with detailed descriptions based on multi-shot videos, weakening their performance on holistic understanding and temporal modeling. For our models, VideoChat2-SUM-shot achieves an overall score of 40.5, surpassing the compared models and MiniGPT4-SUM-shot on all three subtasks. This performance underscores the benefits of video pretraining and the advanced visual backbone of VideoChat2. These baseline results highlight the complexities and demanding nature of our Shot2Story-QA task. We show some example predictions in Figure 5(c). Please refer to Appendix C.2 for more qualitative results.

4 Conclusion

In this work, we present Shot2Story, a large-scale benchmark for comprehensive multi-shot video understanding. We provide detailed shot-level captions for both visual signals and human narrations. Furthermore, we provide comprehensive video summaries based on shot-level captions and design a challenging video question-answering benchmark for multi-shot video understanding. With the rich and diverse descriptions, our benchmark serves as a playground for future multi-modal video understanding models, ready to be extended for a range of other video understanding tasks, such as visual grounding and video-based conversation.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *ICCV*, pp. 5803–5812, 2017.

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. *Advances in Neural Information Processing Systems*, 36, 2024.

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, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

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.

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*, pp. 376–380, 2014.

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.

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)*, pp. 776–780. IEEE, 2017.

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.

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*, pp. 10867–10877, 2023.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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.

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*, pp. 706–715, 2017.

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.

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*, 2023a.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*, 2023c.

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*, 2023d.

Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023e.

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

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

OpenAI. Gpt-4. URL `https://chat.openai.com/?model=gpt-4`.

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*, pp. 311–318, 2002.

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*, pp. 8748–8763. PMLR, 2021.

Enxin Song, Wenhao Chai, Guanhong 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.

Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020.

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.

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*, pp. 4566–4575, 2015.

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.

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

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*, pp. 5036–5045, 2022.

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.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple llm framework for long-range video question-answering. *arXiv preprint arXiv:2312.17235*, 2023.

Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32, 2018.

Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *CVPR*, pp. 6578–6587, 2019.

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.

A ANNOTATION PROCESS

In this section, the annotation process for Shot2Story is detailed, including single-shot caption annotation in Sec. A.1, GPT-4 summary generation prompts in Sec. A.2, human correction of summaries in Sec. A.3, and annotation of question-answering pairs in Sec. A.4.

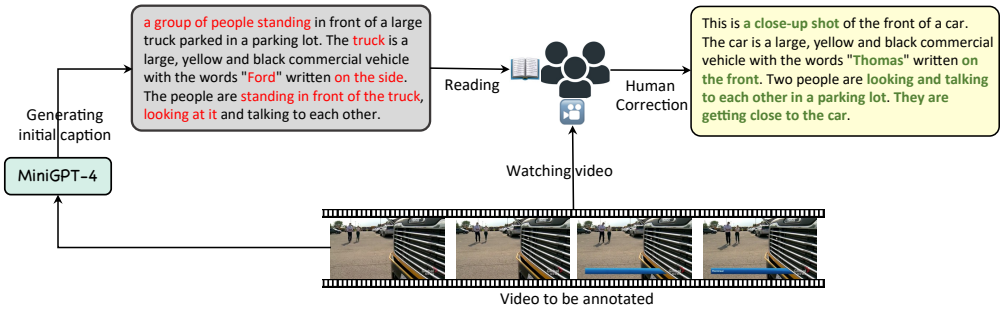

Figure 7: Human annotation process of sing-shot video captions. Texts in bold green represent correct content, while those in red indicate errors. Please find more explanations in Sec. A.1.

A.1 HUMAN ANNOTATION OF SINGLE-SHOT CAPTIONS

Our single-shot video caption annotation process, described in Sec. 2.2, 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 7, 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 during model deployment has been detailed in Sec. 3.2.

A.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 (Li et al., 2023b), which defines video captions and audio captions for each shot, as depicted in Figure 9. For each video, we organize shot durations, video captions, narration captions, and ASR into a text format (see Figure 8). This arranged content is then fed into GPT-4 for generating the video summary.

A.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 10 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.

A.4 ANNOTATION OF VIDEO QUESTION-ANSWERING PAIRS

As shown in Figure 13, our annotation of question-answering (QA) pairs utilizes a hybrid manual-automatic approach to ensure both diversity and quality. Initially, we employ GPT-4 to generate

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

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 annotater 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 9: Prompt template for GPT-4 summarization.

candidate question-answer pairs based on our specific instructions, focusing on creating temporal-related, holistic understanding, and audio-related questions. Subsequently, annotators review these pairs while watching the corresponding videos to eliminate simple or incorrect entries. Subsequently, the annotators are asked to categorize the QA pairs into predefined categories, *i.e.*, temporal-related, holistic understanding and audio-related, and discard the data not under these categories.

The prompts used for generating candidate QA pairs are detailed in Figure 11 and Figure 12. For different task-specific questions, we adopt different task instructions as shown in Figure 11. During generation, we adopt the template shown in Figure 12. The boldfaced texts, such as "**{shot_caps}**", "**{video_sum}**", "**{task_inst}**", are replaced with shot captions organized as in Figure 8, video summary and task-specific instructions from Figure 11.

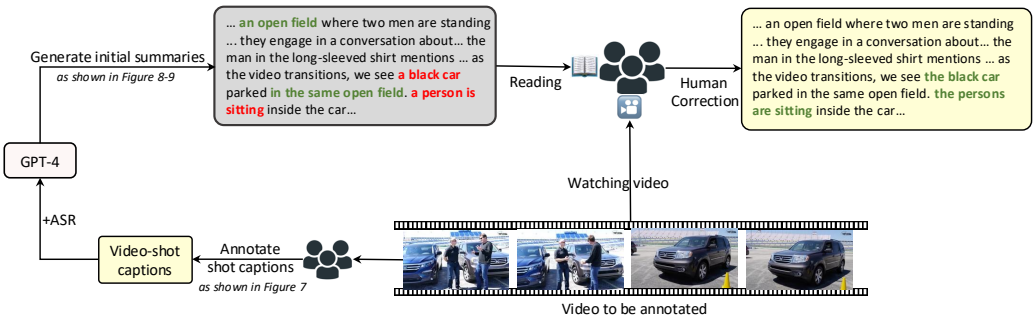

Figure 10: Human correction process of video summaries. Overlapped text is omitted for clarity. Texts in bold green represent correct content, while those in red indicate errors. Please find more explanations in Sec. A.3.

Please ask 5 questions only related to visual content. Please remember to avoid audio content. Please answer the questions according to the Holistic Video Summary, or Video Shot Captions, depending on the question types.

The questions should be related to multiple video shots, rather than a single shot. They can only be answered with multiple video shots and cannot be answered by a single shot. The questions should be logical, reasonable and non-trivial. Please only return json of the questions and answers in the below format.

(a) Temporal-related task instruction

Please ask 5 more new questions only related to visual content. Please remember to avoid audio content. Please answer the questions according to the Holistic Video Summary, or Video Shot Captions, depending on the question types.

The questions should be related to holistic video understanding, such as the topic, the progression and the story line. The questions should be logical, reasonable and non-trivial. The questions should only be answered by viewing and understanding the full video content. Please only return json of the questions and answers in the below format.

(b) Holistic understanding task instruction

Please ask 2 to 5 new questions that can only be answered together with visual and audio contents. Please remember to avoid audio content. Please answer the questions according to the Holistic Video Summary, or Video Shot Captions, depending on the question types.

The questions should be that they can not be answered without audio content. The answer can only be obtained by associating visual content and audio content. Easy questions are strictly prohibited. Remember: Objective questions are always preferred. Please avoid general questions and general answers. General questions are questions that are not specific to the video. General answers mean answers can provided without viewing the video content. An example question is: whom in what appearance speaks what content? This is only for illustration.

(c) Audio-related task instruction

Figure 11: Task instruction for candidate question-answering pairs generation.

B 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 Sec. B.1. For video summarization models like SUM-shot, SUM-holistic, and SUM-text, the prompts are thoroughly explained in Sec. B.2.

B.1 Prompt for single-shot video 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 14. In the figure, boldfaced text, such as "**{asr}**", is replaced with specific video information. The arrangement of visual tokens and text prompts, as presented in Figure 4, is not included here for brevity.

You are a professional video creator. You are required to ask valuable and non-trivial questions according to the given video. Please be notified that the video has multiple shots, containing event progression and storyline. There's also a possibility that the video contains the same objects, subjects and scenes in different video shots.

START VIDEO!

{shot_caps}

The video content in the format of Holistic Video Summary:
{video_sum}

END VIDEO!

START QUESTIONING!
{task_inst}

NOTE:
0. Please ask questions that answers are definite. You can ask questions such as major event, theme or topical category. But you should avoid questions that can be easily answered in logic and content.
1. Please avoid asking repeated questions.
2. Please do not mention the specific video shot index. It is better to describe the starting and ending of a period of content by former and latter actions, events, and objects. Of course, you can use timestamps instead.
3. The questions together should be able to be composited to express the original video content, including event content, event progression, subject appearance and topics.

The questions can definitely be answered by the textual video content. Please generate answers for the questions based on the provided textual video content. The answers should be grounded on the video content, concisely.

Figure 12: Prompt template for candidate question-answering pairs generation.

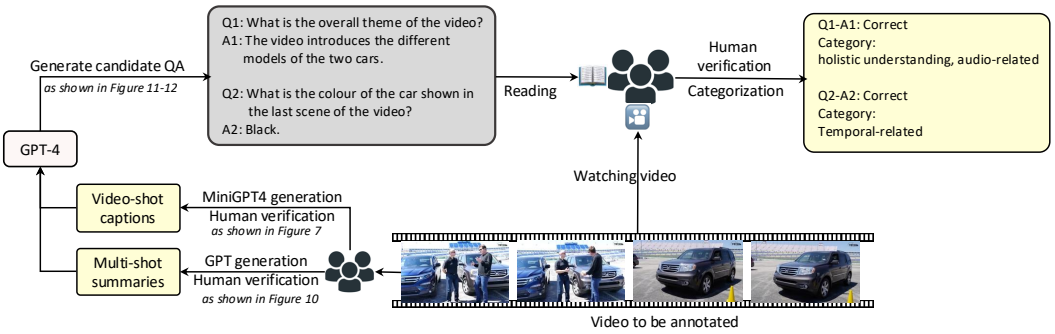

Figure 13: Human annotation process of QA pairs. Please find more explanation in Sec. A.4.

B.2 Prompt for video summarization

In Sec. 3.3, we explore different model variants for video summarization, namely MiniGPT4-SUM-shot, MiniGPT4-SUM-holistic, and VideoChat2-SUM-shot. For training these models, we use the same text prompt as in single-shot video captioning, shown in Figure 14. The key distinction between MiniGPT4-SUM-shot and MiniGPT4-SUM-holistic lies in the arrangement of visual tokens: MiniGPT4-SUM-shot incorporates shot-specific information such as shot number or index along with visual tokens from each shot, whereas MiniGPT4-SUM-holistic uniformly samples frames across the video. For VideoChat2-SUM-shot, the input prompt is the same to MiniGPT4-SUM-shot.

B.3 Prompt for in-domain video question-answering

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 15.

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 14: Prompt during training for single-shot video 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 15: Prompt used for video question answering with summaries.

B.4 Prompts used for zero-shot QA

In this section, we detail the prompts employed for the zero-shot video QA task, which is discussed in Sec. 3.4. Note that we use the same prompt with Vicuna for text summary-based video question-answering, both zero-shot video QA and in-domain video QA.

B.4.1 LLM QA prompt

The text prompt is shown in Figure 15, the same to in-domain video QA. It requires video summaries, which are generated from models trained with our Shot2Story data, which fullfils the definition of zero-shot tasks. To better align with the ground truth answers, we prompt the LLM to generate concise answers solely based on the provided video content.

B.4.2 Evaluation prompt

In our paper, we follow the same evaluation procedure as outlined in (Maaz et al., 2023), using ChatGPT-3.5 to assess the alignment of the generated answers with the given ground truth.

B.5 GPTV generation prompt

The prompt used for GPTV generation for another 90K videos is shown in Figure 16. We follow a similar prompt structure as Figure 8 to organize the different shots. In addition to the shot structure, we additionally embed speaker information and ASR texts for visual-audio correlation in generated summaries. Using Whisper-X, we extract the speaker diarization and ASR texts, which are organized in the form of "*somebody* speaks *something* during *when* to *when*". The speaker diarization descriptions are then appended to the tail of video content in the prompt, in order to have speaker identification information in the generated summaries. We show two samples in Figure 17, which shows the successful speaker assignment to the visual elements and adequate details. Note this subset of data is not verified by human annotators. We only used it in the video summarization experiments in Section 3.3 of the main paper.

C Additional Experiment Results

C.1 Comparison of different models

In this subsection, we present an example of video summarization from our Shot2Story testing split in Figure 18, using MiniGPT4-SUM-holistic, MiniGPT4-SUM-shot, and VideoChat2-SUM-shot. Both MiniGPT4-SUM-shot and VideoChat2-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,

You are a chatbot that conducts conversations based on video contexts. These are frames of a short video. You mainly answer based on the given frames. 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. Do not include details that you are not sure of.
Please note that some speakers in the audio appear in the video, of whom the speech content should be described in the shot. Please also note that some speakers may not appear in the video, who may be background voice or camera holders, of whom the speech content should be described as narrator or background voice.

The video has **{num_shots}** shots. Each shots may contain multiple actions, scenes and subjects.
The **{shot_idx}** shot starts from **{start_duration}**s to **{end_duration}**s.
It contains frames **{frames}**
The ASR contained in this shot is: **{shot_ASR)**
... # more shots in the same structure

{speaker_diarization}

Please create a detailed description that outlines the key actions and components. Please describe the appearance, clothing, and surrounding environment of the characters. Also, please describe the appearance and characteristics of key objects. When multiple people and objects appear, please describe them all and make their descriptions as unique as possible. You should ensure the description is in narrative style and third-person view. You should describe the video coherently

You should describe and include the speech content into each video shot. You should be aware that the speakers in the audio may not appear in the video. If the speakers in the audio does not appear in the video, you should still mention the speech content if it is related to the visual content and topic. The speech content should be properly rephrased from its original ASR texts, if possible. You should ignore the incomplete speech content. Do not mention the words of "ASR". DO NOT include a conclusion of ASR or speakers.
You should not mention duration of videos and video shots. 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. Please describe the video coherently. You should not include a separate conclusion paragraph.

Figure 16: Prompt for video summary generation using GPTV.

and the woman speaking to the camera. VideoChat2-SUM-shot, with its advanced vision backbone and video pretraining, captures more nuanced action details, like "using a wooden spatula to roast almonds in an oven". MiniGPT4-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".

C.2 QA SUMMARY

In this subsection, we present the results of zero-shot video question-answering using Vicuna v0-13B, based on textual summaries of video samples from MSRVTT-QA (Xu et al., 2017) and ActivityNet-QA (Yu et al., 2019). Despite the limitations of our summarization model, which scores 8.6 in CIDEr on the Shot2Story test split (see Table 3), 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 19, `video7089` from MSRVTT-QA portrays a TV show outside the domain of Shot2Story . 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 20, we present the video `v_mZYWfmsYQPA` from ActivityNet-QA. The video's duration is 104 seconds, which is significantly longer than the average duration in our Shot2Story . Our summary effectively identifies important elements such as the main subject's clothing described as "dressed in a black polo shirt with a logo", the actions including "address the

Speaker diarization: There is 1 speaker. SPEAKER_00 speaks during 0.01 seconds to 15.3 seconds.

Human speech: what exactly is going on. So there's still pressure on the clutch pedal. Like the clutch feels like it works. It's just super, super soft. I know it's not a throw bearing clip issue. Like we all know Evos have their pole style clutch. So there's

GPTV-generated video summary:

In the first shot of the video, a man appears to be outside during daylight, with a clear sky above him. He wears a casual, dark hoodie with the letters "ASYN" on it and a grey beanie. He seems puzzled, his brows furrowed as he questions what exactly is happening, which implies confusion or concern regarding a situation off-camera. Transitioning into the second shot, focus shifts to the inside of a vehicle, where the man's hand is interacting with the clutch pedal, pressing it down and articulating it to test its feel. He expresses that while the clutch pedal still has pressure, it feels unusually soft. The footwell of the vehicle is shown, highlighting the clutch pedal as a key object of interest. We see the hand pressing and wiggling the clutch pedal, conveying an examination of its functionality. In the final shot, the narrative continues with a further detailed explanation about the car's mechanical state. There's an assertion that it is not an issue with the throw bearing clip, a component known to be problematic in certain car models, which are alluded to be "evos" known for their pull-style clutches. The focus remains on the vehicle interior and briefly moves to the exterior, suggesting the man has stepped out, possibly inspecting other parts of the car relevant to the clutch issue. The final frame ends with the man's hand gesturing, emphasizing his point about the car's mechanical specifics, possibly indicating the clutch's pole style.

Speaker diarization: There is 1 speaker. SPEAKER_00 speaks during 0.8 seconds to 21.9 seconds.

Human speech: Here's what you'll need. Beef rump roast, Roma tomatoes, yellow onion, garlic, oil, a little cilantro, some salt and pepper, and last but certainly not least, Herde's Red Guajillo Chile Mexican cooking sauce and roasted pasilla chili sauce. To start, heat oil over medium-high heat. Now take your beef and season

GPTV-generated video summary:

In the first shot of the video, a woman dressed in a professional chef's coat is standing in a kitchen, holding a large dish that appears to be filled with a prepared meal, possibly a salad. She is poised in front of a kitchen counter adorned with various ingredients and cooking utensils. The kitchen features modern appliances and a decorative painting on the wall. As she stands there, an off-screen speaker begins with, "Here's what you'll need," suggesting that the viewer will be instructed on a recipe or cooking method. Transitioning to the second shot, the focus shifts to the ingredients and equipment required for the recipe. Several frames show a close-up of a bowl filled with cut beef rump roast, which is the primary ingredient. The kitchen counter is neatly arranged with Roma tomatoes, a diced yellow onion, minced garlic, and a clear bottle of oil. Fresh cilantro, salt, pepper, and two jars of Herdez brand cooking sauces, one labeled as Guajillo Chile and the other as Roasted Pasilla Chile, are also displayed. A list of ingredients appears on the screen, detailing the quantities and specific items needed, including "2 ½ pounds beef round rump roast, cut into 1-1½ inch cubes" and the other components mentioned. The speaker narrates the list of ingredients, including the beef rump roast, Roma tomatoes, yellow onion, garlic, oil, cilantro, salt and pepper, as well as the specific Herdez Mexican cooking sauces. The video transitions once more into the third shot, where the cooking process begins. A pot is centered on a stove, and the chef pours oil into it. Following this, the chef seasons the beef with salt, preparing it for cooking. The action implies that the beef will soon be browned in the oil. The speaker's instructions -- to heat oil over medium-high heat and to season the beef with salt -- align with the visual actions, demonstrating the initial steps in preparing the dish. The viewer is now both informed of the ingredients and shown the beginnings of the cooking process, paving the way for the continuation of the recipe..

Figure 17: Samples of video summaries generated by GPTV. In both cases, GPTV successfully correlates the audio speaker and the person in the video, *i.e.* “He expresses that” and “an off-screen speaker begins”. Moreover, GPTV-generated summaries capture the overall action and event flow with the help of shot structure, and the essential details in the video.

audiance” and “demonstrate the proper grip”, and the setting, noted as “a ping pong table in an indoor sports facility”. 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, four 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 Q2, Q5, Q6 and Q9.

Shot2Story-QA: In Figure 21, we present a video sample from our Shot2Story-QA dataset. The questions address audio-related (Q1, Q4), holistic understanding (Q2, Q3), and temporal-related (Q4) aspects. With shot structure enabled in VideoChat2-SUM-shot, the generated video summary includes actions and transitions, such as *begins with a close-up shot of a person holding a pedal*, *transitions to a man*, *then shifts to a wider shot*, *returns to a close-up shot of the man*, and *concludes with a wider shot of the man*. However, the generated summary falls short in capturing detailed information about the actions and their orders (e.g., "Taking out a piece of paper" in Q3), and fails to align visual and audio content (e.g., different components in Q1 which have been mentioned in ASR texts). As a result, Vicuna struggles to answer these questions, highlighting the challenges posed by our Shot2Story-QA benchmark. To further improve the model performance, the summarization

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.

MiniGPT4-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.

MiniGPT4-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.

VideoChat2-SUM-shot video summarization:

in the video, a woman with blonde hair, wearing a light blue top, is standing in a kitchen with wooden cabinets in the background. she is holding a glass bowl filled with almonds and a blender. as she talks to the camera, she mentions that there are a couple of tricks that people need to know in order to be successful. the scene then transitions to a close-up shot of a person using a wooden spatula to roast almonds in an oven. the woman's voice can be heard in the background, explaining that roasting the almonds helps to bring out not only the flavors but also the oil. finally, the video returns to the woman in the kitchen, who continues to hold the glass bowl filled with almonds and the blender.

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 18: Example for video summarization using MiniGPT4-SUM-holistic, MiniGPT4-SUM-shot and VideoChat2-SUM-shot. All three models effectively grasp the video's main topic and content. Please find explanations in Sec. C.1.

video7089.mp4

VideoChat2-SUM-shot video summarization:

in the first shot of the video, three individuals are seated at a desk in what appears to be a television studio with a cityscape view in the background. the man on the left is wearing a dark shirt, the woman on the right is dressed in a light blue blouse, and the man on the right is wearing a black t-shirt. they are engaged in a lively conversation, with the man on the left gesturing with his hands as he speaks. the woman on the right seems to be listening attentively, while the man on the right listens as well. the conversation seems to revolve around a topic related to the city, with the man on the left expressing his enthusiasm for it. transitioning to the second shot, the scene shifts to a stage where two men are standing. the man on the left is dressed in a light blue blazer and a dark shirt, while the man on the right is wearing a dark shirt and khaki pants. the man on the left is holding a ukulele, and they both appear to be in the middle of a performance or presentation. the man on the left is smiling and seems to be enjoying the moment, while the man on the right stands with his hands on his hips, looking at the audience or perhaps the man with the ukulele. in the third and final shot, the video returns to the studio setting with the three individuals from the first shot. the man on the left is still wearing his dark shirt, the woman on the right is in her light blue blouse, and the man on the right is in his black t-shirt. the man on the left is gesturing with his hands again as he speaks, while the woman on the right and the man on the right listen attentively. the woman on the right seems to be responding to something the man on the left is saying, and the man on the right listen intently.

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 19: Example for video question-answering using summaries - MSRVTT. We show the zero-shot video summary using our trained VideoChat2-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 Sec. C.2.

v_mZYWfmsYQPA.mp4 *duration: 104 seconds*

VideoChat2-SUM-shot video summarization:

in the opening scene of the video, a man stands on a ping pong table in an indoor sports facility. he is dressed in a black polo shirt with a logo on the left chest, and he holds a ping pong racket in his right hand. the background is a vibrant blue, and there's a sign that reads welcome to the world of ping pong. as he speaks, he introduces the topic of the video, which is about the basics of ping pong. transitioning to the second scene, the same man continues to address the audience, now holding a red ping pong ball in his hand. he emphasizes the importance of understanding the basics of ping pong, suggesting that this is a fundamental aspect of the sport. in the third scene, the man is seen again on the ping pong table, this time preparing for a serve. he demonstrates the proper grip and stance for a serve, which is a crucial part of the game. the video then cuts to the fourth scene, where a different man appears on the ping pong court. he is wearing a white t-shirt emblazoned with the website www.pingponginfo.com and the text ping pong information . this man speaks about the importance of proper footwork in ping pong, highlighting the role it plays in the game. finally, in the fifth and last scene, the same man from the first scene reappears. he is still on the ping pong table, and he poses a question to the viewers, asking if they have a question about ping pong. this serves as an invitation for the audience to engage with the content and potentially learn more about the sport. throughout the video, the speakers, including the man on the ping pong table and the man in the white t-shirt, provide informative commentary on the basics of ping pong, emphasizing the importance of understanding the fundamentals of the game.

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 demonstrating the basics of ping pong, including proper grip, stance, serve, and footwork. ✅
A2: No, 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, the person is standing on a ping pong table in an indoor sports facility. ✅
A5: Based on the video, it is not clear what happened before the man started playing table tennis. ❌
A6: The video is taken in an indoor sports facility with a vibrant blue background and a sign that reads Welcome to the world of ping pong. ❌
A7: Based on the video, ping pong is a safe sport. The video shows people of different ages and skill levels playing the game without any major incidents, and the speakers emphasize the importance of safety in the sport, such as proper grip and stance for serves. ✅
A8: The benefit of this exercise is to learn the basics of ping pong, including proper grip and stance for a serve, the importance of footwork in the game, and the importance of understanding the fundamentals of the sport. By watching the video, viewers can learn about the basics of ping pong and potentially gain a deeper appreciation for the sport. ✅
A9: There are two people in the video. ❌
A10: The person in the video wears a black polo shirt with a logo on the left chest. ✅

Figure 20: Example for video question-answering using summaries - ActivityNet. We show the zero-shot video summary using our trained VideoChat2-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 Sec. C.2.

Human speech: We get pedals disc for the front wheel, im assuming screws, probably for the disc bell brackets skewer and an instruction book hang on.

VideoChat2-SUM-shot video summarization:

the video begins with a close-up shot of a person holding a pedal in their hand. the scene then transitions to a man in a black t-shirt standing in a garage filled with various items, including a bicycle and a motorcycle. he is engaged in a task at a workbench, where a bicycle is mounted on a stand. as he works, he mentions that he is using a pedal disc for the front wheel. the video then shifts to a wider shot of the man in the black t-shirt, still standing in the garage filled with items. he continues his task at the workbench, where a bicycle is mounted on a stand. the scene then returns to a close-up shot of the man in the black t-shirt, still working at the workbench. the video concludes with a wider shot of the man in the black t-shirt, still standing in the garage filled with items. he continues his task at the workbench, where a bicycle is mounted on a stand.

Questions:

Q1: What are the different components the man handles from the beginning to the end of the video?
Q2: How does the setting change from the start to the end of the video?
Q3: What is the man's immediate action after handling the skewer?
Q4: What is the main activity being depicted in the video?

Answers:

GT1: A pedal, disc brake pad, screws, bell, brackets, skewer, paper, and instruction book.
GT2: The setting remains the same, in the garage with the half-finished bike.
GT3: Taking out a piece of paper.
GT4: The items for the bike assembly or preparation.

Vicuna answers using summaries & GPT evaluation:

A1: The different components the man handles from the beginning to the end of the video are: a piece of equipment, a white object, and the pedals disc for the front wheel. ❌
A2: The setting changes from a workshop or laboratory to a garage. ❌
A3: The man is placing the skewer back into the box. ❌
A4: The main activity being depicted in the video is the demonstration of tool preparation for bikes in a garage setting. ✅

Figure 21: Example for video question-answering using summaries - Shot2Story-QA. We show the in-domain video summary using our trained VideoChat2-SUM-shot model. The summary can capture the shot transitions and major actions. However, as indicated by the QA results, the detailed actions and their order fail to match with the groundtruth, such as whether the immediate action after handing the skewer. Additionally, it fails to provide adequate information regarding the consistent scene between shot transitions, such as “at a workbench” and “in the garage”. Please find explanations in Sec. C.2.

model could be enhanced with detailed information with denser frames, fine-grained matching of audio and visual cues, and tracking of the same objects across the video. We leave it as future work.

D ADDITIONAL IMPLEMENTATION DETAILS

Video-Shot Captioning. For each video shot, we uniformly sample 4 frames. For testing, a consistent text prompt is used, *i.e.*, "The audio transcripts are asr. Describe this video in detail.". The maximum number of new tokens generated by the LLM is capped at 180 for both training and inference.

MiniGPT4-SUM-shot, MiniGPT4-Holistic and VideoChat2-SUM-shot. In two SUM-shot models, 4 frames per video shot are sampled uniformly. In MiniGPT4-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, *i.e.*, "The audio transcripts are {asr}. Describe this video in detail.", is used during inference.

For video-shot captioning and video summarization tasks, both models are trained on 8×2 A100-80G GPUs using Pytorch. The captioning model is trained for 10 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 Sec. B.

VAST. We tune the model following the official instructions by inputting our video frames, audios and ASR texts. The optimizing target is the concatenation of our single-shot visual caption and single-shot narration caption. During inference, we input the corresponding modalities, in case of two different model versions we trained (visual + audio, or visual + audio + ASR texts), into the model and use the directly generated captions for evaluation. The model is trained on 8×4 A100-80G GPUs for 3 epochs, 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. 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 MSRVTT-QA (Xu et al., 2017) and ActivityNet-QA (Yu et al., 2019) 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 15. This integrated content is then processed through Vicuna (Chiang et al., 2023) to obtain answers. The evaluation of these results is carried out following the methodology outlined in (Maaz et al., 2023).

E BROADER IMPACT

Data Limitations and Ethical Considerations. We provide cropped multi-shot videos instead of the original videos. Users can also turn to download these from original sources. Given HD-VILA-100M (Xue et al., 2022)'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. Specifically, our videos don't have personally identifiable information or offensive content, which is ensured by the manual annotation process. The authors will take the responsibility of long-term maintenance.

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 (Touvron et al., 2023) 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 in advancing language modelling and multi-modal learning.

Future Research and Development. We 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 8×2 A100-80G GPUs, we are aware this may be a limitation. Consequently, we plan to focus future 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.