SALOVA: Segment-Augmented Long Video Assistant for Targeted Retrieval and Routing in Long-Form Video Analysis

Supplementary Material

A. Details of SceneWalk Dataset

A.1. Detailed Data Statistics

We provide a comprehensive analysis of the proposed SceneWalk dataset, focusing on detailed data statistics, including video duration, categorical distribution, and segment-level descriptions. The information emphasizes the versatility and diversity of the dataset, ensuring its applicability for training our video-LLM.

Dataset Composition. The SceneWalk dataset comprises 87,867 long-form video sources, spanning a total of 11.87 Khrs (average video duration: 486.5 seconds). The video sources are collected from a curated selection of 10 diverse categories as in Fig. 1 sourced primarily from YouTube, with additional contributions from Movie & Drama datasets [21, 53]. This ensures a wide range of real-world scenarios, avoiding static categories.

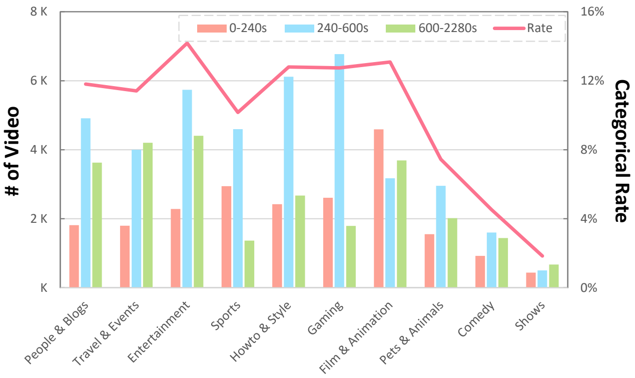


Figure 4: Detailed video duration range statistics for each video category in the SceneWalk dataset.

Video Duration Distribution. The collected videos can be split into three distinct duration ranges to analyze temporal diversity: (i) 0–240 seconds (short): This range constitutes about 24.4% of all segments, (ii) 240–600 seconds (long): This intermediate range accounts for the largest proportion, approximately 46.1% of the dataset, and (iii) 600–2280 seconds (extreme-long): The longest duration range comprises around 29.5% of the dataset. The distribution of video durations for each video category is illustrated in Fig. 1(b) (outer circle), and more detailed duration distributions can be found in Fig. 4.

A.2. Pipeline for Dense Caption

Splitting into Video Segments To divide untrimmed and long video sources into a massive 1.29M video segments, we directly utilize PySceneDetect with the AdaptiveDetector using the default adaptive threshold (3.0), which compares the difference in content between adjacent frames similar using a rolling average of adjacent frame changes. This can help mitigate false detections in situations such as fast camera motions.

Instructions of Dense Segment Captioning. To generate detailed descriptions for each video segment obtained from the above process, we mainly use a pre-trained LMM (VILA-1.5-13B [36]). Below Tab. 4 includes the instructions for generating those captions. We randomly select one from the list and use it as a query for the model.

Video Caption Generation Instruction:

- “Provide a detailed description of both the visual content and the storyline depicted in the video.”
- “Thoroughly describe the scenes, actions, and characters featured in the video”
- “Elaborate on the visual and narrative elements of the video in detail.”
- “Describe what is happening in the video in detail, including both the visual details and the narrative context.”
- “Describe every element of the video, from visual details to the unfolding narrative, including how each aspect interacts to enhance the storytelling.”
- “Offer a granular analysis of the video, detailing the scenes, character actions, and dialogue, alongside any symbolic visual elements that add depth to the story.”
- “Narrate the unfolding events in the video with attention to both the visual composition and the plot, describing how each scene visually portrays the narrative tensions or themes.”

Table 4: The list of instructions for detailed description for each video segment.

Captioning and Scoring. Each segment is densely captioned, generating highly detailed textual descriptions that average 137.5 words per segment. Please see the densely captioned video examples in Fig. 6 and Fig. 7. To ensure alignment quality, a generalized bipartite matching framework is employed: (i) Video-to-Text (V2T) Correspon-

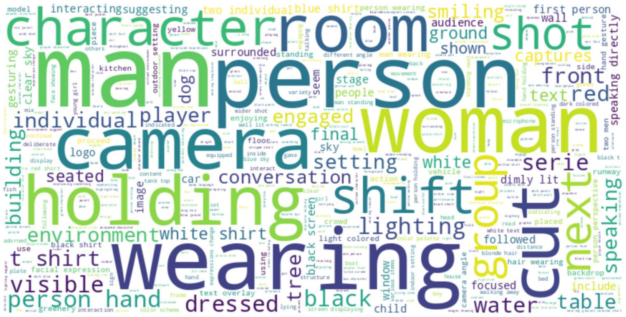


Figure 5: WordCloud analysis of the SceneWalk dataset.

dence: A matrix evaluates the alignment between video segments and their paired captions using LanguageBind [73], and (ii) Text-to-Text (T2T) Context Similarity: The textual coherence among adjacent captions is assessed using SBERT [54], enhancing overall alignment robustness.

Supervision from Correspondence Scores. To derive the supervision signal y_i for training, we leverage the correspondence scores S_{V2T} (Video-to-Text) and S_{T2T} (Text-to-Text), as discussed in Sec. 4. For each correspondence score matrix, we apply thresholding to extract meaningful relationships. Specifically, we define thresholds τ_{V2T} and τ_{T2T} for the two matrices, and elements with scores exceeding these thresholds are treated as positive correspondences (th: 0.18 (τ_{V2T}) and 0.8 (τ_{T2T}), respectively). These positive correspondences are then one-hot encoded to form binary matrices Y_{V2T} and Y_{T2T}, where each element indicates whether a specific correspondence is valid. Finally, we compute the union of these binary matrices to produce the final supervision signal (*e.g.,* $y_i{=}Y_{\text{V2T}} \cup Y_{\text{T2T}}$). The union operation ensures that any correspondence deemed valid by either of the two modalities contributes to the final supervision. This approach captures both the multi-modal alignment (Video-to-Text) and intra-modal coherence (Text-to-Text), providing a robust supervision signal for the retrieval task.

The resulting y_i is then incorporated into the similarity loss function $\mathcal{L}_{\text{sim}}$ as described in Eq. (1), ensuring that the model effectively learns the nuanced relationships between video segments and their corresponding textual descriptions. By combining S_{V2T} and S_{T2T} in this manner, we account for the complexity of generalized bipartite matching and enhance the model's ability to align correspondences across and within modalities.

A.3. Word Cloud Analysis.

The Word Cloud visualization in Fig. 5 highlights the richness and diversity of visual cues captured within the SceneWalk dataset. The prominent keywords such as *man*, *woman*, *person*, and *group* reflect the dataset's strong emphasis on human-centric descriptions, focusing on capturing the presence, actions, and interactions of individuals within a scene. These terms highlight that the dataset prioritizes detailed portrayals of people as central subjects, providing rich context about their appearance, activities, and relationships with the surrounding environment.

Furthermore, the inclusion of descriptive spatial and contextual terms (*e.g., stage*, *floor*, *light*, *tree*, *etc*,.) illustrates how the dataset prioritizes capturing environmental details alongside subject interactions. This level of granularity ensures that the visual-textual mappings are comprehensive, enabling the dataset to serve as a robust resource for training models that require an in-depth understanding of scene composition and narrative continuity.

By focusing on such fine-grained visual details, the SceneWalk dataset can provide generic scene descriptions, encapsulating nuanced visual content that is critical for multimodal tasks. The highlighted terms reflect not only the dataset's diversity but also its deliberate emphasis on actionable visual semantics, making it particularly valuable for an intermediate training step, as proposed in Sec. 4.2, by enabling models to effectively learn and represent long video knowledge, including scene comprehension and nuanced understanding.

B. Training Details of SALOVA

Training Config. In this section, we elaborate on the training process of SALOVA-7B. All variations of SALOVA are trained under unified settings, except for the choice of visual encoders. Note, however, that *per-device batch sizes* may vary slightly due to hardware limitations. To equalize the *global batch size* across these variations, *gradient accumulation* is implemented, facilitating a consistent training timeline for each variant. The detailed training configuration for each step can be found in Tab. 5, which optimizes the use of available GPU memory for batch sizing and en-

config	Stage1	Stage1.5	Stage2
input modality	image, video	video	video
input frame		1 FPS	
input resolution		384 × 384	
optimizer		AdamW ($\beta_1, \beta_2{=}0.9, 0.999$)	
lr schedule		cosine decay	
training precision		BFloat16	
DeepSpeed train		ZeRO-2	
warmup epochs		0.03	
trainable params	connectors	full	full
lr_{vision, text}	-	2e-6	2e-6
lr_{LLM, others}	1e-3	2e-5	2e-5
global batch size	256	8	64
total epochs	1	1	1
Max token drop	0.0	0.7	0.4

Table 5: Training hyper-parameters for different stages. Here, connectors indicates SR-Router and ST-Connector.

sures efficient training dynamics with limited hardware resource.

C. Architecture Details of SALOVA

Network Config. Here, we explain our network configurations in detail. For the first part of our architecture, the Spatio-Temporal Connector, we employ the Perceiver Resampler [2] architecture (but, smaller size), which consists of a 2-layer, 2-head Transformer architecture followed by a 2-layer MLP with GELU activation as a projector. For the connector's latent features, we set the number of latent features to 256 and the hidden size to 1024. Next, the second module, the Segment Retrieval Router, consists of a 2-layer, single head Transformer architecture. The Transformer uses a d_model of 1024 and PReLU as the activation function.

D. Additional Experiments

Results of LongVideoBench. Due to the page limit of the main manuscript, in this additional section, we elaborate on both validation and test set results of LongVideoBench [57] for further demonstration. As in Tab. 6, it shows an analogous tendency for Video-MME benchmark, which exhibits a significant performance increase after the short duration video (15s∼). We highlight again that such trend is mainly due to the retrieval capability of SALOVA, which excels in associating visual content with contextual information, even as video lengths increase.

		LongVideoBench					
Model	Size	8-15s	15-60s	180-600s	900-3600s	test set	val set
Proprietary LMMs							
GPT-4o [46]	-	71.6	76.8	66.7	61.6	66.7	66.7
Gemini 1.5 Pro [50]	-	70.2	75.3	65.0	59.1	64.4	64.0
GPT-4-Turbo [44]	-	66.4	71.1	61.7	54.5	60.7	59.1
Open-sourced LMMs							
VideoChat2 [32]	7B	38.1	40.5	33.5	33.6	35.1	36.0
VideoLLaVA [35]	8B	43.1	44.6	36.4	34.4	37.6	39.1
PLLaVA [59]	7B	45.3	47.3	38.5	35.2	39.2	40.2
LLaVA-1.5 [37]	7B	45.0	47.4	40.1	37.0	40.4	40.3
ShareGPT4Video [8]	7B	**46.9**	50.1	40.0	38.7	41.8	39.7
Ours							
SALOVA-3B	3B	46.3	46.7	41.9	39.8	42.2	41.4
SALOVA-3.8B	3.8B	45.3	48.3	42.6	40.6	42.9	41.6
SALOVA-7B†	7B	46.0	**50.7**	**44.4**	**42.1**	**44.5**	**43.5**

Table 6: Comparison results for LongVideoBench. The best results are highlighted in **bold** and the runner-up results are underlined. Note that the † mark indicates our efficient size for the frontier model utilizing CLIP [48] as vision encoders (smaller resolution and 144 visual tokens per frame).

Ablation Study for Retrieval Number. In addition, we conduct an analysis of the number of video segments used for inference on the Video-MME benchmark. In our architectural design, the number of video segments can be dynamically set based on retrieval estimates from the SR-Router, which forwards partial yet pertinent spatio-temporal information from the video to the LMMs. We compare how varying the number of video segments affects performance results. Note that the maximum number of video segments in Video-MME is 13. As in Tab. 7, we clearly observe that increasing the number of frames tends to enhance performance. However, performance saturates after the retrieval number reaches 9. This saturation may be related to the fact that excessive input information becomes more disruptive than helpful for reasoning about partial scenes in the video.

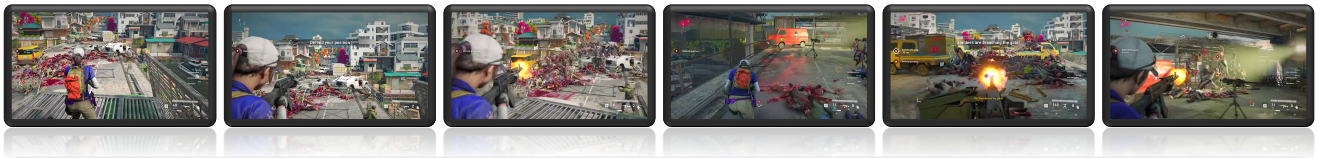
Ablation	Video-MME			
	Short: $\leq$2m	Mid: 4-15m	Long: 30-60m	Overall
Top-k	: Number of Video Segments for Retrieval			
1	48.1	44.4	39.1	43.9
5	48.1	45.0	39.2	44.1
9	48.3	46.3	41.1	45.3
13	48.1	44.7	39.7	44.1

Table 7: Ablation studies on retrieval number for video segments. We utilize SALOVA-3B model with the CLIP [48] (`clip-vit-large-patch14-336`) for computational efficiency.

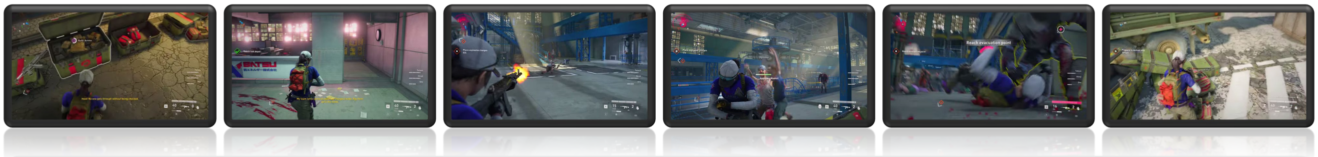
Qualitative Results. We provide qualitative results with varying video lengths to clearly demonstrate the effectiveness of SALOVA across short, medium, and long videos as in Fig. 8. For example, a short video in Fig. 8, our model accurately retrieved a scene of the Moon colliding with the Earth, depicting an astronomical disaster based on the given question. Similarly, in medium-length videos, SALOVA effectively identified the scene where a male judge selects a card corresponding to the question. Specifically, even in videos longer than 40 minutes, SALOVA accurately identifies scenes related to the correct answer, such as people eating BBQ after exploring the history of the food's origin, based solely on the query input and the video content.

These consistent qualitative results across all lengths indicate that the successful retrieval of pertinent video segments relevant to the input query significantly contributes to the model's efficiency. As demonstrated in our analysis, SALOVA effectively handles different amounts of video data, which supports robust scene understanding and reasoning.

Video Source: 15M 11S



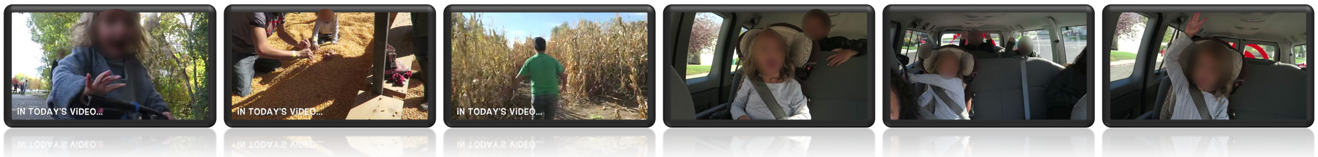
Video Segment (1): [00:00:00.000 - 00:01:24.706]
Description (1): The video opens with a view of a character standing on a rooftop, overlooking a cityscape with buildings and vehicles. The character is equipped with a firearm and is surrounded by a chaotic scene of destruction, with numerous zombies scattered around the area. The character moves through the environment, engaging in combat with the zombies, using a firearm to shoot at them. The zombies are seen in various states of decay, with some being on fire. The character's movements are swift and precise, aiming and firing at the zombies with accuracy. The environment is detailed, with realistic textures and lighting that enhance the immersive experience. The video game's user interface is visible, displaying information such as the number of zombies killed, the remaining time, and the character's health status.



Video Segment (2): [00:01:24.706 - 00:11:26.554]
Description (2): The video is a compilation of gameplay footage from a first-person shooter game. The player character is seen in various combat situations, including shooting at enemies, picking up items, and navigating through different environments. The environments vary from indoor settings with industrial equipment to outdoor areas with vehicles and structures. The player character is equipped with a firearm and is seen aiming and shooting at enemies, who are also armed. The game's HUD (heads-up display) is visible, showing information such as ammo count, health status, and objectives. The video also includes moments where the player character is knocked down and is seen recovering.

⋮

Video Source: 9M 43S



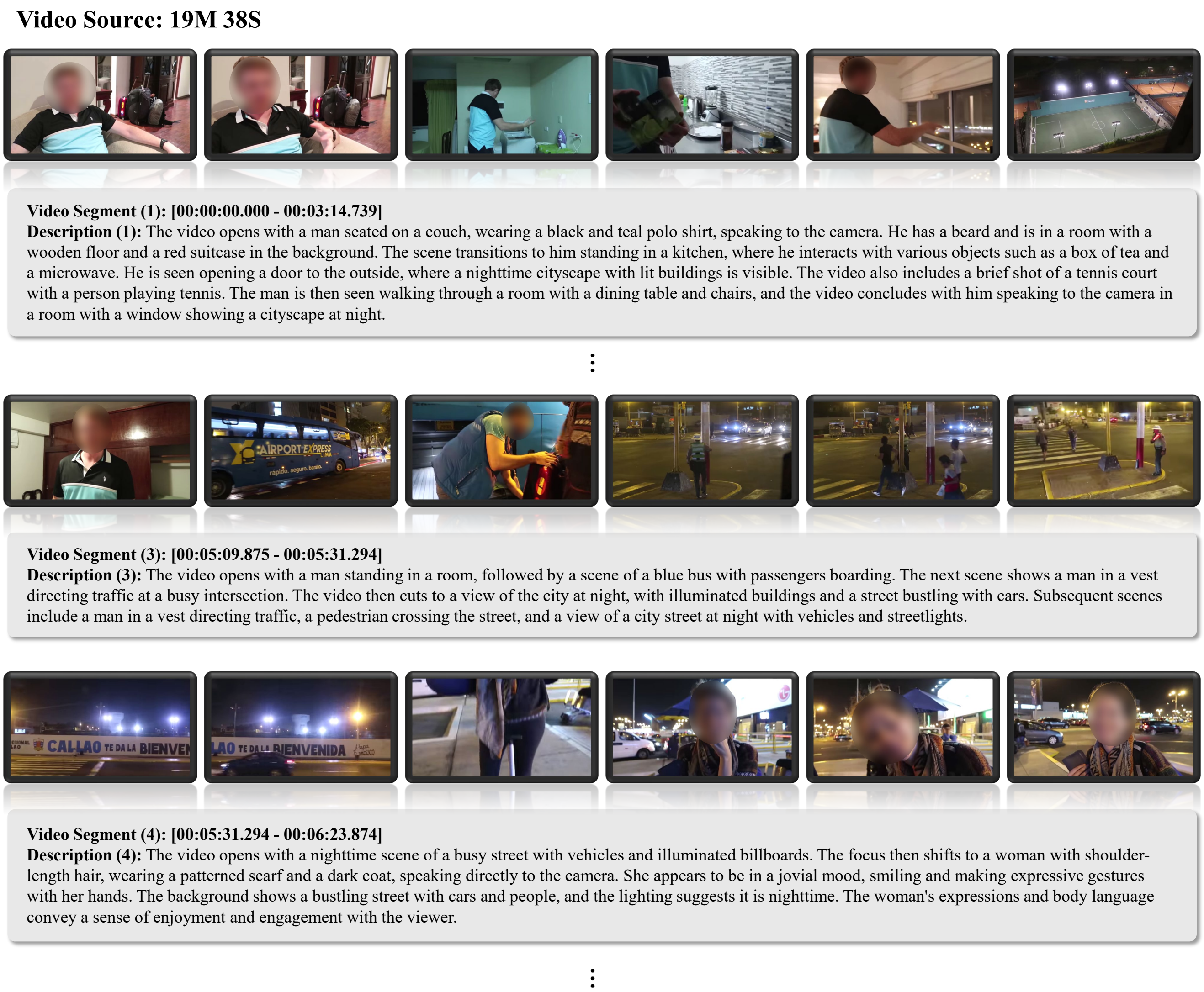
Video Segment (1): [00:00:00.000 - 00:00:26.856]
Description (1): The video opens with a scene of a young girl riding a bicycle on a path with fallen leaves, followed by a scene of her playing in a cornfield. The next scene shows the girl in a car, wearing a seatbelt and smiling. The video then cuts to a scene of the girl and a boy in the back of a car, with the girl in the front seat and the boy in the back seat. The boy is seen waving and smiling at the camera. The video concludes with the girl in the front seat, smiling and waving at the camera.

⋮

Video Segment (4): [00:01:14.964 - 00:01:39.051]
Description (4): The video opens with a scene of a young child riding a red tricycle in a parking lot, surrounded by parked cars and trees with autumn foliage. The child is wearing a light-colored top and dark pants. The next scene transitions to a park where two children are seen playing with a large pile of autumn leaves, tossing them into the air. The children are dressed in casual outdoor clothing suitable for the season. The following scenes depict more children, some on skateboards and others on bicycles, navigating a pathway covered with fallen leaves. The children are wearing jackets and helmets, indicating a cool weather. The video concludes with a shot of the pathway leading through a wooded area with trees displaying autumn colors, suggesting a peaceful and leisurely atmosphere.

⋮

Figure 6: Examples of the SceneWalk dataset (i).



Video Source: 19M 38S

Video Segment (1): [00:00:00.000 - 00:03:14.739]
Description (1): The video opens with a man seated on a couch, wearing a black and teal polo shirt, speaking to the camera. He has a beard and is in a room with a wooden floor and a red suitcase in the background. The scene transitions to him standing in a kitchen, where he interacts with various objects such as a box of tea and a microwave. He is seen opening a door to the outside, where a nighttime cityscape with lit buildings is visible. The video also includes a brief shot of a tennis court with a person playing tennis. The man is then seen walking through a room with a dining table and chairs, and the video concludes with him speaking to the camera in a room with a window showing a cityscape at night.

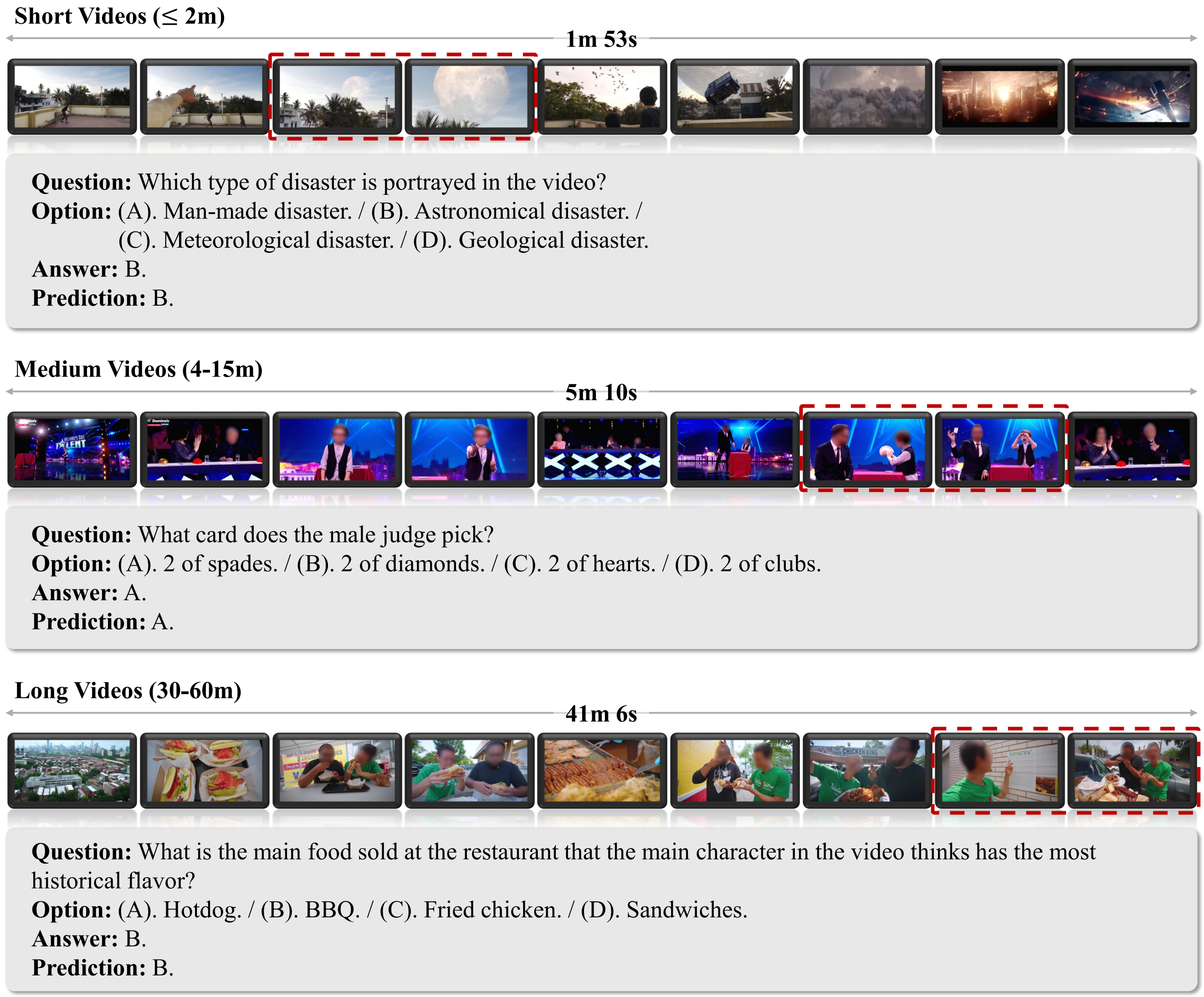
⋮

Video Segment (3): [00:05:09.875 - 00:05:31.294]
Description (3): The video opens with a man standing in a room, followed by a scene of a blue bus with passengers boarding. The next scene shows a man in a vest directing traffic at a busy intersection. The video then cuts to a view of the city at night, with illuminated buildings and a street bustling with cars. Subsequent scenes include a man in a vest directing traffic, a pedestrian crossing the street, and a view of a city street at night with vehicles and streetlights.

Video Segment (4): [00:05:31.294 - 00:06:23.874]
Description (4): The video opens with a nighttime scene of a busy street with vehicles and illuminated billboards. The focus then shifts to a woman with shoulder-length hair, wearing a patterned scarf and a dark coat, speaking directly to the camera. She appears to be in a jovial mood, smiling and making expressive gestures with her hands. The background shows a bustling street with cars and people, and the lighting suggests it is nighttime. The woman's expressions and body language convey a sense of enjoyment and engagement with the viewer.

⋮

Figure 7: Examples of the SceneWalk dataset (ii).



Question: Which type of disaster is portrayed in the video?
Option: (A). Man-made disaster. / (B). Astronomical disaster. /
(C). Meteorological disaster. / (D). Geological disaster.
Answer: B.
Prediction: B.

Question: What card does the male judge pick?
Option: (A). 2 of spades. / (B). 2 of diamonds. / (C). 2 of hearts. / (D). 2 of clubs.
Answer: A.
Prediction: A.

Question: What is the main food sold at the restaurant that the main character in the video thinks has the most historical flavor?
Option: (A). Hotdog. / (B). BBQ. / (C). Fried chicken. / (D). Sandwiches.
Answer: B.
Prediction: B.

Figure 8: Qualitative examples on Video-MME [20] with SALOVA-7B. Note that the red dashed lines indicates the top-1 relevant video segment estimation for the question.