ReWind: Understanding Long Videos with Instructed Learnable Memory

Anxhelo Diko[1*†] Tinghuai Wang[2*] Wassim Swaileh[2] Shiyan Sun[2] Ioannis Patras[2]
[1] La Sapienza University of Roma
[2] Huawei Helsinki Research Center
diko@di.uniroma1.it
{tinghuaiwang,shiyansun,wassim.swaileh,ioannis.patras}@huawei.com

Model	LLM	Backbone	MSRVTT-QA		MSVD-QA	
			Accuracy	Score	Accuracy	Score
Short-Video Approaches						
Video Chat [12]	Vicuna-7B	ViT-G	45.0	2.5	56.3	2.8
Video LLaMA [17]	Vicuna-7B	ViT-G	29.6	1.8	51.6	2.5
Video-ChatGPT [14]	Vicuna-7B	ViT-L	49.3	2.8	64.9	3.3
LLaMA Adapter [18]	LLaMA-7B	ViT-L	43.8	2.7	54.9	3.1
Chat-UniVi [10]	Vicuna1.5-7B	ViT-L	55.0	3.1	69.3	<u>3.7</u>
Long-Video Approaches						
MA-LMM [8]	Vicuna-7B	ViT-G	48.5	-	60.6	-
VTimeLLM [9]	Vicuna1.5-7B	ViT-L	-	-	-	-
MovieChat [15]	LLaMA2-7B	ViT-G	52.7	2.6	**75.2**	**3.8**
LLaMA-VID [13]	Vicuna-7B	ViT-G	**58.3**	<u>3.3</u>	<u>69.7</u>	<u>3.7</u>
ReWind (Ours)	LLaMA2-7B	ViT-G	<u>56.9</u>	**3.4**	69.3	**3.8**

Table 1. Evaluation for short VQA on MSRVTT-QA, MSVD-QA, and VideoChatGPT test sets with GPT-3.5. The best result is highlighted in **bold**, and the second best is <u>underlined</u>.

1. Additional Short-Video VQA Experiments

In Table 1, we report zero-shot results on MSRVTT-QA and MSVD-QA of ReWind. Our approach, despite being specifically designed for long videos, outperforms previous short-term video approaches and has competitive performance compared to the current SOTA. Specifically, we obtain SOTA results on *Score* for both datasets.

2. Pretraining Stage

During the pretraining stage, we conduct standard multimodal alignment, to train the perceiver component to effectively capture semantic visual information. Specifically, as reported in the main manuscript, it consists of contrastive learning utilizing the SigLIP [16] loss between perceiver projections and caption encodings from BERT [5]. Particularly, we use a pre-trained BERT from hugging face under the following repository name: '*google-bert/bert-base uncased*'[1]. In this stage, the ViT (EVA02-ViT-G/14) is kept frozen and the features are extracted from its penultimate layer as suggested in [6]. As reported in the manuscript, the pertaining is done on 100K video-caption pairs randomly selected from the WebVid2.5M and Panda70M datasets.

Dataset	#Clips	Avg. Duration (sec.)	Avg. Text len
WebVid2.5M [2]	2.5M	18	12
Panda-70M [4]	70.8M	8	13.2

Table 2. Comparison of WebVid2.5M and Panda-70M Datasets.

WebVid2.5M. WebVid2.5M is a large-scale video-text dataset comprising 2.5 million video-text pairs predominantly sourced from YouTube [2]. It encompasses diverse topics and genres, offering a representative sample of real-world video content. This dataset is instrumental for large-scale training of models for tasks such as video captioning, retrieval, question answering, and video-language pretraining. Notably, WebVid2.5M employs weakly labeled data, where text annotations are extracted from sources like titles and descriptions, presenting a realistic challenge for video-language models.

Panda70M. Panda-70M is a large-scale video-caption dataset. It consists of 70 million high-quality video-caption pairs, derived by splitting 3.8 million long videos from the HD-VILA-100M dataset into semantically coherent clips [4]. Multiple cross-modality teacher models were employed to generate diverse captions for each clip, with a fine-grained retrieval model subsequently selecting the most relevant caption as the final annotation. Panda-70M is intended for large-scale training in tasks such as video captioning, video and text retrieval, and text-driven video generation.

Additionally, as mentioned in the manuscript, this stage involves 10K training steps with a batch size of 64, using the AdamW optimizer and cosine scheduling. The learning rate is set to 1e-4 with 500 warmup steps. The entire training process is done in half-precision on 8 GPUs. For each video, in this stage, we uniformly sample 64 frames and resize the frames to 224×224.

*Equal Contribution
†Work done while at Huawei
[1] https://huggingface.co/google-bert/bert-base-uncased

3. Instruction Tuning

The second training stage engages also the memory module, the DFS, and the LLM (LLaMA-2, from huggingface: '*"meta-llama/Llama-2-7b-chat-hf"*[2] with LoRA. This phase employs the instruction-tuning strategy on on ∼100K samples from the VideoChatGPT [12] dataset together with the previous 100K samples used for pertaining, aiming to integrate all network components seamlessly for the VQA.

VideoChatGPT. The Video-ChatGPT dataset comprises approximately 100K video-instruction pairs designed to enhance multimodal conversational AI models [12]. The annotations combine human-assisted efforts and semi-automatic methods. In the human-assisted stage, expert humans are used to enrich the details of the existing ActivityNet [3] dataset. On the other hand, during the semi-automatic stage, using advanced vision-language models like BLIP-v2 [11] and GRIT enables scalable and high-quality labeling of spatial, temporal, and contextual video content.

During this stage, as reported in the main paper ReWind is trained for 100K steps with a batch size of 64, a learning rate of 5e-5, and 2K warmup steps, using the AdamW optimizer. LoRa is configured with a rank of 64 and an alpha of 32 and quantized in 4-bit during training. The training is done on 8 V100 GPUs. Additionally, the frames are selected with a constant frame rate of 1fps and are not bound to a fixed number of frames.

For the temporal grounding task, ReWind is further trained with instruction tuning on DiDemo and ActivityNet datasets for an extra 15K steps. During this training stage, we use 500 warmup steps and a batch size 64.

DiDeMo. The DiDeMo [1] dataset is a large and diverse benchmark designed for temporally localizing events in videos based on natural language descriptions. It consists of videos collected from Flickr, each trimmed to a maximum duration of 30 seconds. These videos are segmented into 5-second intervals to simplify the annotation process. The dataset includes a total of 26,892 moments, with each moment possibly linked to multiple textual descriptions, offering detailed accounts that often specify camera movements, temporal transitions, and activities.

ActivityNet. The ActivityNet [3] dataset is a large-scale video benchmark for temporal action localization, captioning, and VQA. It consists of approximately 20,000 untrimmed videos sourced from YouTube, covering 200 different activity classes. Each video typically contains an average of 1.41 annotated activities, with temporal boundaries provided for precise action localization.

[2] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

4. Evaluation Datasets.

We evaluate our model on three datasets: MovieChat1K, Charades-STA, and VideoChatGPT dataset. Here, we provide information regarding MovieChat1K and Charades-STA. For VideoChatGPT, please refer to Section 3.

MovieChat1K. The MovieChat1K [15] dataset is a benchmark for evaluating video understanding through long-form video question answering. It comprises 1K long videos (9.13 minutes on average) annotated with 14K manually created questions and answers. The dataset is tailored to assess models' capabilities in processing extensive temporal contexts and understanding complex visual narratives in videos. It supports two VQA modalities: global and breakpoint. The global setting requires processing the entire video and answering questions about its content, while the breakpoint mode involves processing the video up to a specific timestamp and answering questions about the event at that point. Each video has one dense description, 3 global questions, and 10 breakpoint questions with timestamps.

Charades-STA. The Charades-STA dataset is designed for temporal activity localization in videos based on natural language queries. It is derived from the Charades [7] dataset and includes temporal annotations linking video clips to textual descriptions. Each query is paired with a specific start and end time within a video, allowing models to learn to locate activities matching the query. The dataset consists of 12,408 video-sentence pairs for training and 3,720 pairs for testing, with videos featuring everyday indoor activities (157 activity categories).

5. Additional Qualitative Esamples

Figures 1, 2, and 3 present further qualitative evaluations of ReWind, complementing the results within the manuscript. These figures illustrate three distinct cases: First, Figure 1 demonstrates ReWind's capabilities in a cooking scenario, showcasing its ability to describe video content, temporally localize actions, and highlight frames selected via DFS. Second, Figure 2 evaluates ReWind's VQA proficiency in long videos using an 8-minute video from the MovieChat1K test set. Finally, Figure 3 assesses ReWind's VQA performance on a video from the VideoChatGPT validation set.

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Int. Conf. Comput. Vis.*, pages 5803–5812, 2017. 2

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Int. Conf. Comput. Vis.*, pages 1–8, 2021". 1

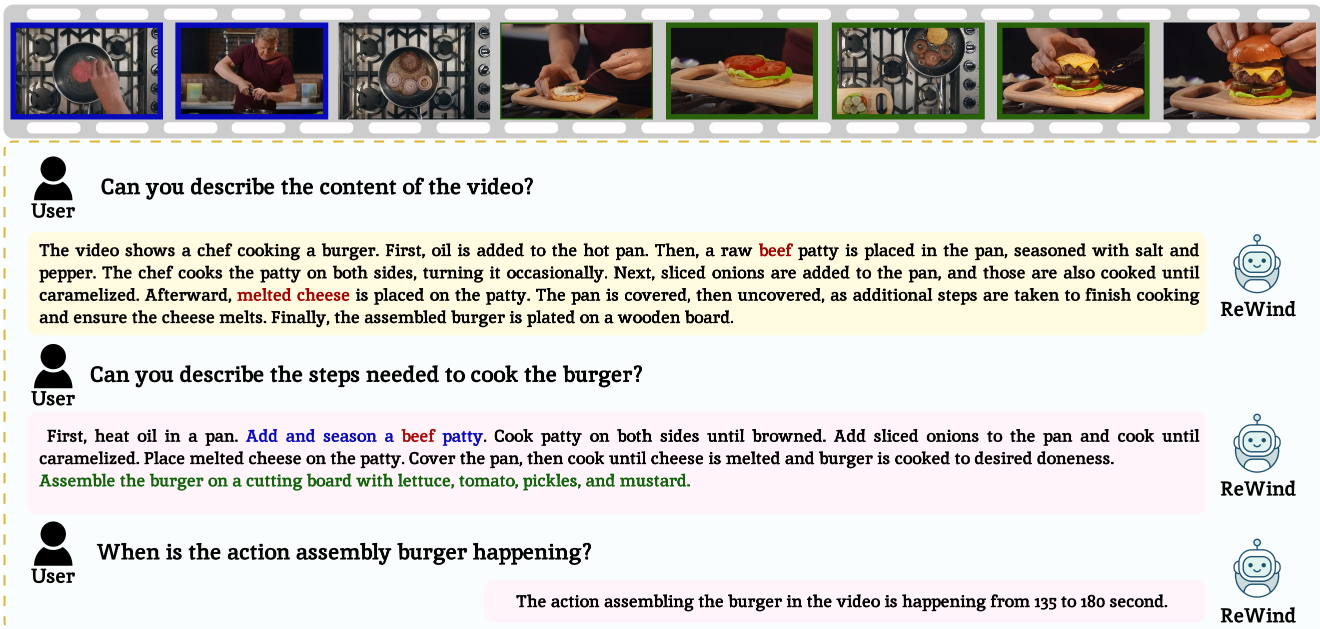

Figure 1. Qualitative evaluation using a cooking video over 3 minutes long, ReWind successfully described both the overall video content and the specific steps performed, demonstrating accurate temporal localization of events. To highlight potential areas of improvement, details that may be hallucinated based on language priors are indicated in red text. Highlighted frames were selected using the DFS mechanism, with text color corresponding to the frame to easily connect descriptive details to their visual counterparts in the video.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Int. Conf. Comput. Vis.*, pages 961–970, 2015. 2

[4] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13320–13331, 2024. 1

[5] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, pages 1–10, 2018. 1

[6] 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. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19358–19369, 2023. 1

[7] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Int. Conf. Comput. Vis.*, pages 5267–5275, 2017. 2

[8] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13504–13514, 2024. 1

[9] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 14271–14280, 2024. 1

[10] Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13700–13710, 2024. 1

[11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Int. Conf. Mach. Learn.*, pages 19730–19742. PMLR, 2023. 2

[12] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1–8, 2023. 1, 2

[13] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *Eur. Conf. Comput. Vis.*, pages 1–14, 2024. 1

[14] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Association for Computational Linguistics*, pages 12585–12602, 2024. 1

[15] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 18221–18232, 2024. 1, 2

[16] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Int. Conf. Comput. Vis.*, pages 11975–11986, 2023. 1

[17] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Conf. on Empirical Methods in Nat. Lang. Process.*, pages 543–553, 2023. 1

[18] Renrui Zhang, Jiaming Han, Chris Liu, Aojun Zhou, Pan Lu, Yu Qiao, Hongsheng Li, and Peng Gao. Llama-adapter: Efficient fine-tuning of large language models with zero-initialized attention. In *Int. Conf. Learn. Represent.*, pages 1–10, 2024. 1

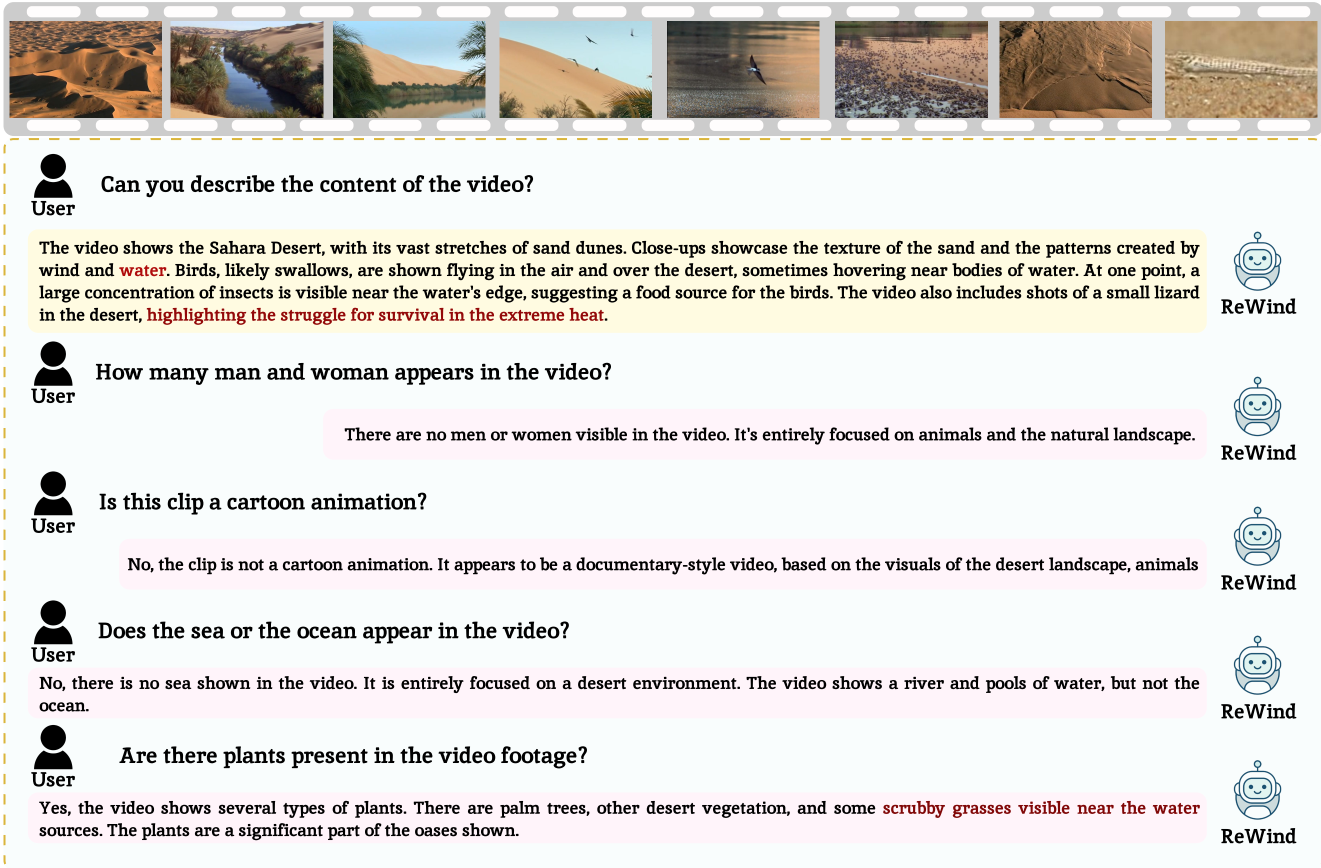

Figure 2. Qualitative evaluation on an +8 minutes video from MovieChat1K test set. ReWind is able to process this long video and answer questions of different natures, including a detailed description. Details hallucinated based on language priors are indicated in red text

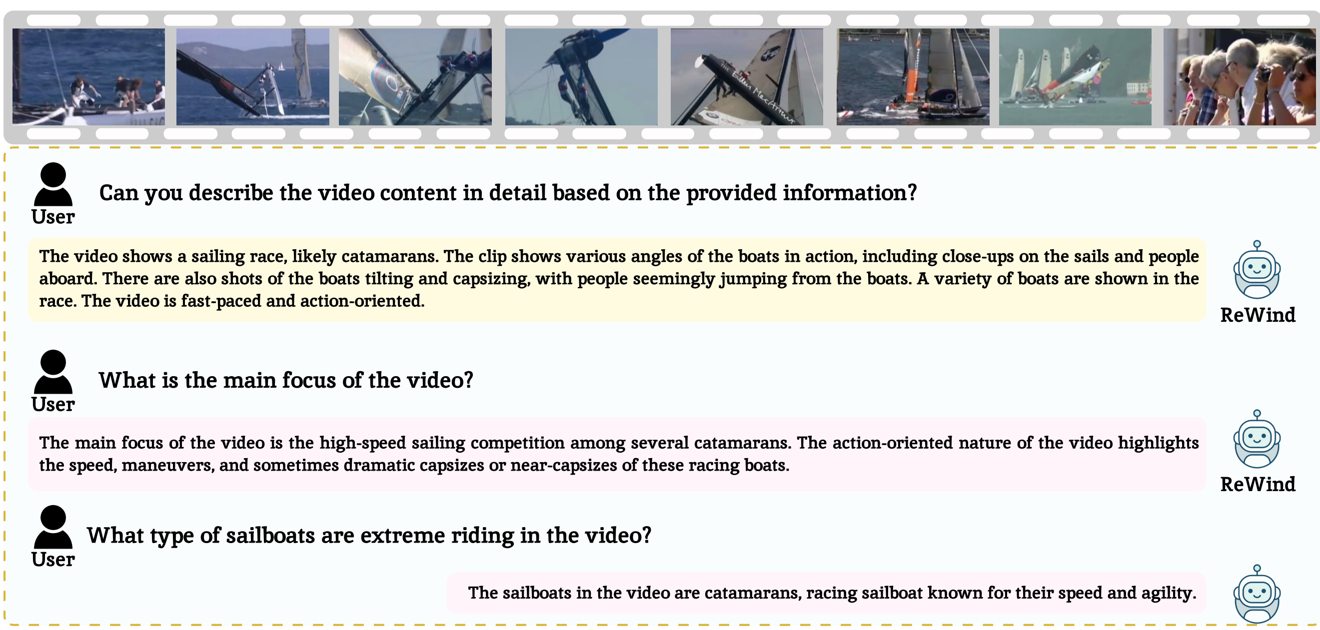

Figure 3. Qualitative example on a +3 minutes video from VideoChatGPT dataset. ReWind is able to correctly extract necessary information and answer to questions.