LiveCC: Learning Video LLM with Streaming Speech Transcription at Scale

Supplementary Material

1. LiveCC-7B Demo

This section showcases four demo videos[1] to demonstrate the capability of our LiveCC-7B to provide real-time commentary in real-world videos across different domains, including sports (football), science (astronomy), news (weather forecast), and instructional (computer repair) videos. As illustrated in Figure 3-6, in the first demo, our LiveCC-7B model correctly recognizes all exact penalty timings, highlighting its strong temporal perception abilities. By leveraging the extensive world knowledge gained from watching millions of YouTube videos, our model accurately reports the name of the related player. The second demo showcases the model's ability to comment beyond sports by precisely presenting astronomy knowledge and demonstrating good OCR capability to read large numbers. The third demo further reveals its fine-grained temporal understanding capability, as evidenced by its real-time commentary on subtle changes in weather maps. The final demo demonstrates that our model is also capable of generating a tutorial to guide users, revealing its potential to serve as a real-time assistant.

2. Implemetation Details

2.1. Prompt Template

In this section, we introduce the prompts used during the pre-training, instruction-tuning, and inference stages. As illustrated in Figure 1(a) and (b), the previously transcribed ASR texts are provided as context for the CM task if they are available. Otherwise, we provide the video title as the context. During loss calculation, these context tokens are masked. For the training sequence, we first append visual tokens for every two frames, followed by the timestamp-aligned transcriptions. For the QA task illustrated in Figure 1(c), we follow the format of LLaVA-Video [11] to present the visual tokens of all frames at one time, followed by questions and answers. As for inference, we follow the SFT CM format and remove the commentary tokens, leaving the model to generate them in a real-time manner. For QA tasks, we follow the SFT QA format but remove the answer tokens, which are generated by the model.

[1] The audio in the demo videos are implemented by ChatTTS [1].

2.2. Win Rate Computation on Sports-3K

In this section, we present the detailed process for computing the win rate on Sports-3K-CC. To start, we categorize the models into two groups based on their inference schemes: (**i**) **Clip-wise caption models**, including GPT-4o [4], Gemini-1.5-Pro [2], VideoLLaMA2 [3], LongVA-7B [10], IXC-2.5-7B [9], LLaVA-OV-7/72B [5], LLaVA-Video-7/72B [12], Qwen2VL-7/72B-Instruct [7], Oryx-7B [6]. (**ii**) **Frame-wise streaming model**, *i.e.*, our proposed LiveCC-7B.

For clip-wise caption models, we directly input the overall event clips, perform a **one-time** inference, and use the generated response as the commentary. To ensure stylistic consistency and fair evaluation, the same context as that shown in Figure 1 is applied across all models. Given that LLaVA-Video-72B [12] is the open-source state-of-the-art model on multiple QA benchmarks, its commentary serves as the baseline for comparison with other models. For our LiveCC-7B, we adopt **streaming** inference, where commentary is generated frame by frame. The model leverages both the context and the previously generated content as historical input for future token generation. The generated tokens are then concatenated to form the complete commentary, which is subsequently evaluated for quality.

For evaluation, we prompt GPT-4o-mini [4] to assess whether a given commentary surpasses that

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
[VISUAL TOKENS of 2 frames]
[VIDEO TITLE] (if PREV. ASR not exists)
[PREV. ASR] (if exists)<|im_end|>
<|im_start|>assistant
[WORDS] ...<|im_end|>
<|im_start|>user
[VISUAL TOKENS of 2 frames]
<|im_start|>assistant
[WORDS] ...<|im_end|>
<|im_start|>user
[VISUAL TOKENS of 2 frames]
<|im_start|>assistant
[WORDS] ...<|im_end|>
...
```

(a) Pre-training CM

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
0.0-3.0s
[VISUAL TOKENS of 6 frames]
[PROMPT]
[PREV. ASR] (if exists)<|im_end|>
<|im_start|>assistant
[WORDS] ...<|im_end|>
<|im_start|>user
3.0-4.0s
[VISUAL TOKENS of 2 frames]
<|im_start|>assistant
[WORDS] ...<|im_end|>
<|im_start|>user
4.0-5.0s
[VISUAL TOKENS of 2 frames]
...
```

(b) SFT CM

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
[VISUAl TOKENS OF ALL FRAMES]
Question: [QUESTION]
[OPTIONS]
Please select the correct answer.<|im_end|>
<|im_start|>assistant
Answer: [ANSWER]<|im_end|>
```

(c) SFT QA

Figure 1. The prompts used during the pre-training instruction-tuning (aka. SFT) stages. CM represents commentary, QA denotes question-answering. For pre-training and instruction-tuning, the previous ASR texts are concatenated to form the context for the live commentary task if they are available. Otherwise, the context is formed by the video title. These contexts are masked during loss calculation. Note that QA data is incorporated exclusively during the instruction-tuning stage. As for inference, we remove the groundtruth in the prompts, *i.e.*, the words followed by a frame or the answer to a multiple-choice question.

of LLaVA-Video-72B. The evaluation is based on two key criteria: (**i**) Semantic Alignment, *i.e.*, consider which text conveys the same meaning, details, and key points as the groundtruth ASR transcript, with minimal deviation. (**ii**) Stylistic Consistency, *i.e.*, assesses which text maintains a tone, word choice, and structure similar to the groundtruth transcript. The overall prompt is written as:

```
You will review two generated texts (Text
↪ A and Text B) and compare them to a
↪ ground-truth ASR transcript. Your
↪ task is to select the generated text
↪ that best aligns with the
↪ ground-truth transcript in terms of
↪ both semantic accuracy and stylistic
↪ consistency. Specifically:

1. Semantic Alignment: Consider which
↪ text conveys the same meaning,
↪ details, and key points as the
↪ ground-truth ASR transcript, with
↪ minimal deviation.

2. Stylistic Consistency: Assess which
↪ text maintains a tone, word choice,
↪ and structure similar to the
↪ ground-truth transcript.

Based on these criteria, choose the
↪ generated text that better aligns
↪ with the ground-truth ASR transcript
↪ overall. Your response should only
↪ contain a letter in [A, B] that
↪ indicates your choice.'

Ground-truth ASR transcript: [GT ASR]

Text A: [LLAVA-VIDEO-72B TEXT]

Text B: [MODEL TEXT]
```

The final win rate is calculated as the proportion of instances where GPT-4o-mini selects the model's response over the baseline.

2.3. Response Parsing in QA evaluation

As described in the Section "Experiments", we follow the approach outlined in LMMs-eval [8] to parse the LLM's response into a concrete option during QA evaluation. Our parsing rules are straightforward: (**i**) If the response is an isolated letter indicating the option, it is directly accepted as the answer. (**ii**) If the response does not explicitly

```python
def parse_pred(pred, options, GPT):
    pred = pred.strip()
    if pred.startswith('A.') or pred.startswith('A ') or pred == "A":
        return 'A'
    if pred.startswith('B.') or pred.startswith('B ') or pred == "B":
        return 'B'
    if pred.startswith('C.') or pred.startswith('C ') or pred == "C":
        return 'C'
    if pred.startswith('D.') or pred.startswith('D ') or pred == "D":
        return 'D'

    prompt = (
        'You will be given four options [A,B,C,D] and a sentence describing a
        ↪  choice of in these options. '
        'Please respond with an upper-case letter indicating the option
        ↪  selected by the sentence. '
        'If there are no options match, respond with an upper-case \'E\'.\n'
        '{options[A]}\n'
        '{options[B]}\n'
        '{options[C]}\n'
        '{options[D]}\n'
        'Sentence: {pred}\n'
        'Do not respond with any additional text.'
    )

    return GPT(prompt)
```

Listing 1. The pseudo-code for response parsing in QA evaluation. “pred” denotes the model‘s prediction.

indicate a choice, we use GPT-4o-mini to map the response to the semantically aligned option. The pseudo-code and detailed prompts are provided in Listing 1. From our observations, only Gemini-1.5-pro [2] requires GPT-based parsing, as other models consistently return their choices directly.

3. Additional Experiments

3.1. Response Latency

To highlight the efficiency of our streaming model, we present the response latency of LLaVA-Video-7B/72B alongside our model in Table 1. Response latency is defined as the time a user waits to see the model's output, a critical factor affecting user experience. Since the LLaVA-Video series are trained in a captioning style, requiring a full clip as input rather than a single frame, their response latency is significantly higher than that of our model. Notably, LiveCC not only achieves lower latency but also delivers high-quality commentary, This promising result further reinforces the effectiveness of our proposed dense interleave training paradigm.

Model	Latency	Input	Inf. Type
LLaVA-Video-72B [12]	20.51s	Clip	Captioning
LLaVA-Video-7B [12]	5.62s	Clip	Captioning
LiveCC-7B	**0.36s**	Frame	Streaming

Table 1. The response latency comparison between LLaVA-Video-7/72B and our LiveCC-7B. Inf. is short for Inference.

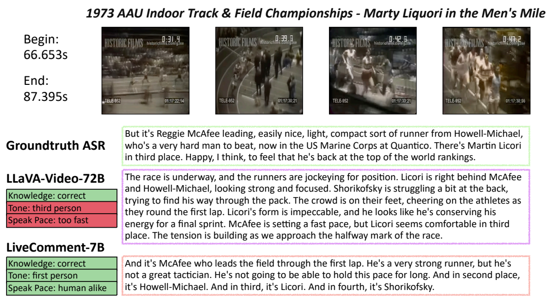
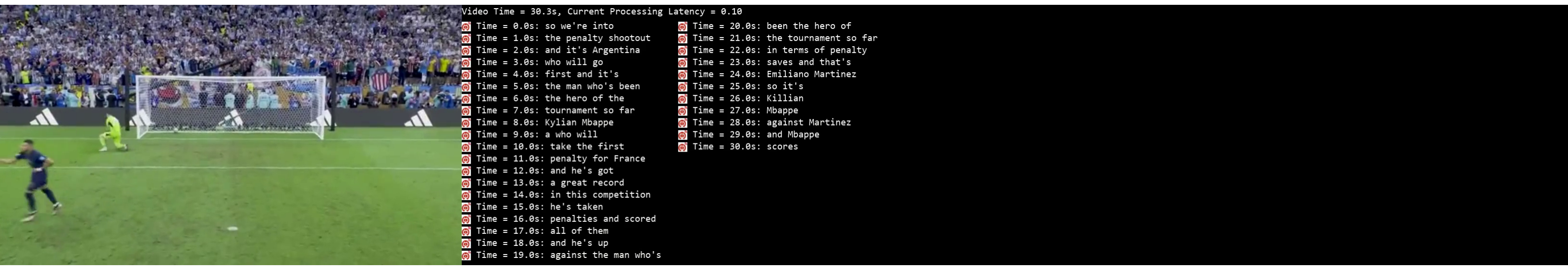


Figure 2. The comparison between the commentary generated by LLaVA-Video-72B and our LiveCommnet-7B.

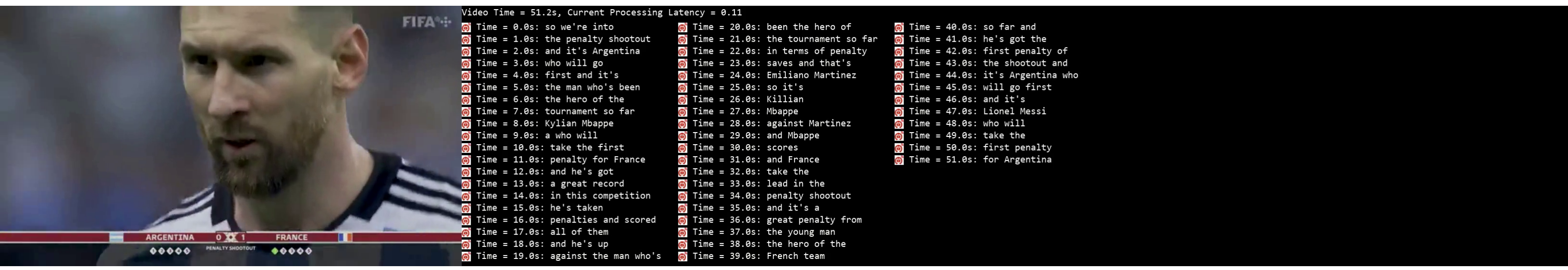
3.2. Commentary Quality

We analyzed the quality of the generated content, as shown in Figure 2. Benefiting from training on millions of ASR-transcribed videos, our model produces commentary that is more aligned with human preferences in terms of tone and speaking pace, while maintaining accurate event understanding. In contrast, the LLaVA-Video-72B, although capable of correctly describing the event, falls short in emulating human-like commentary.

(a) Video Time: 8.6s



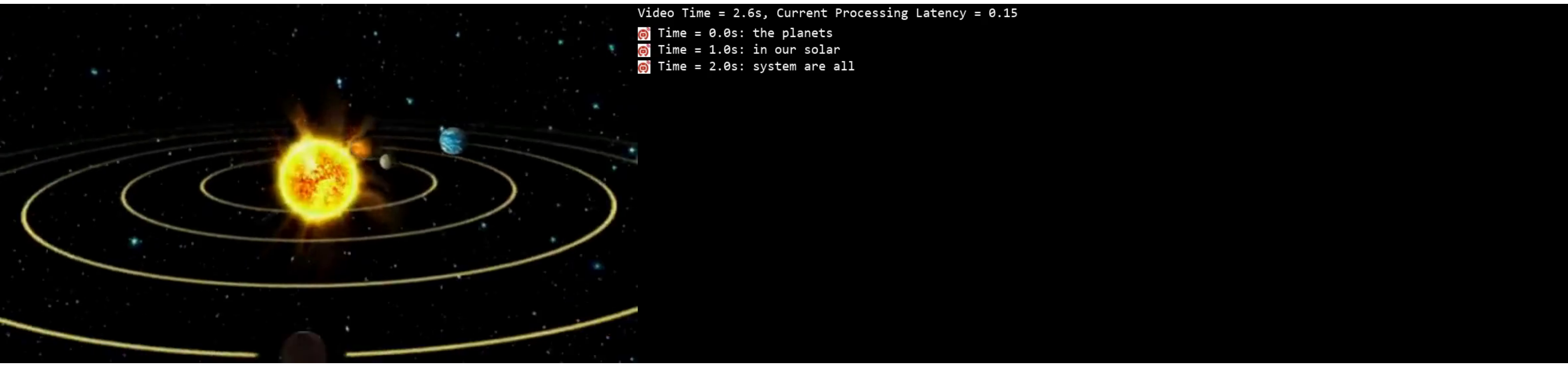
(b) Video Time: 30.3s



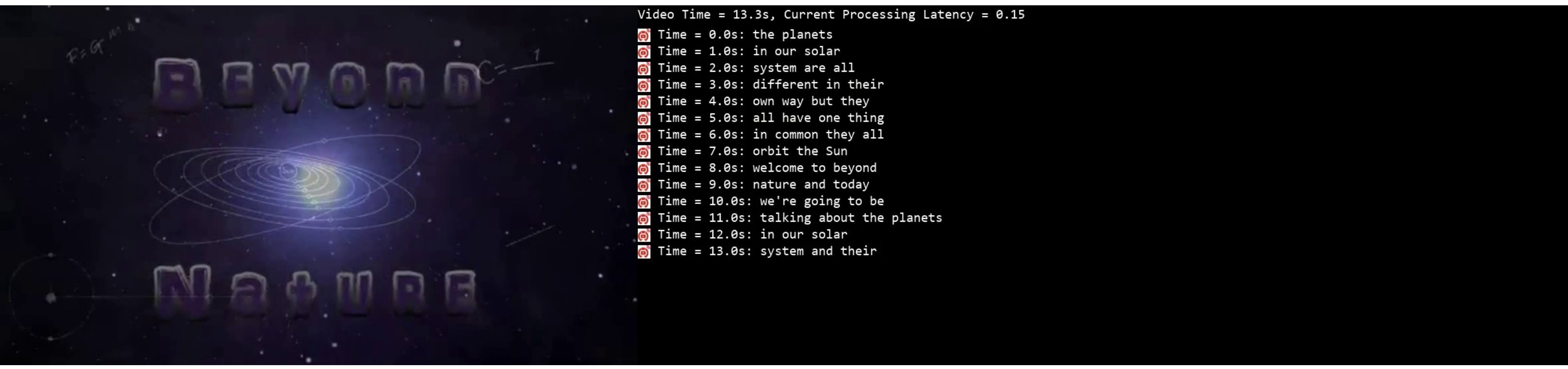
(c) Video Time: 51.2s

(d) Video Time: 77.2s

Figure 3. Real-time video commentary demo on unseen YouTube video (`MCWJNOfJoSM`). The original YouTube title is "Argentina v France: Full Penalty Shoot-out — 2022 FIFAWorldCup Final". We only give a part of YouTube title "Full Penalty Shoot-out — 2022 FIFAWorldCup Final" as prompt to avoid information leakage.



(a) Video Time: 2.6s

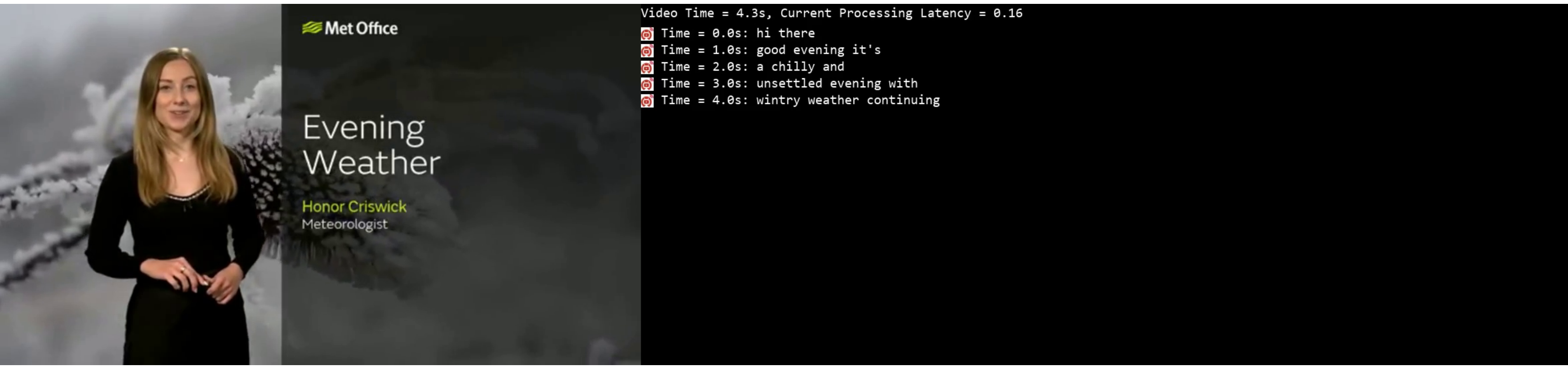


(b) Video Time: 13.3s

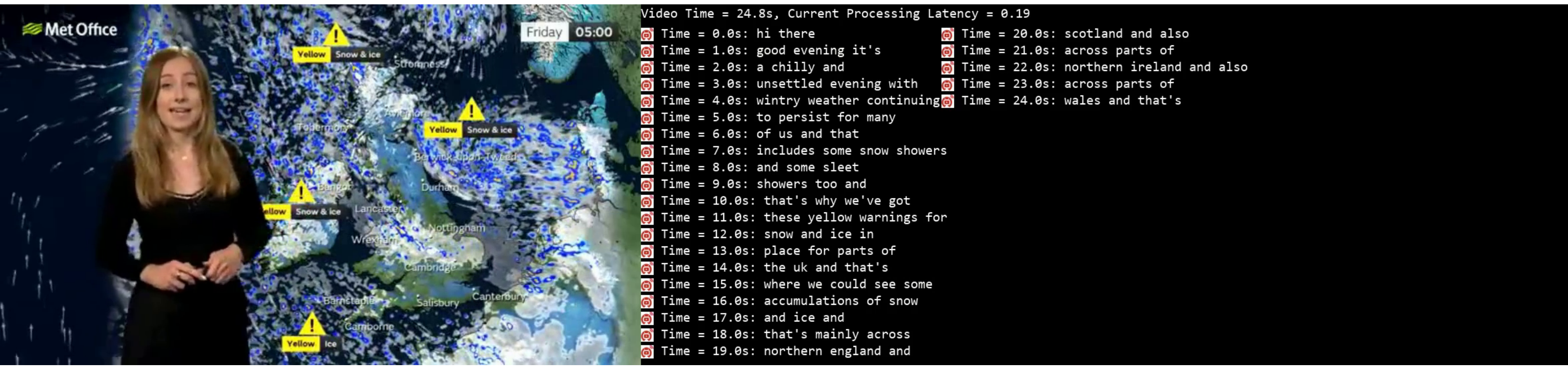
(c) Video Time: 31.6s

(d) Video Time: 45.0s

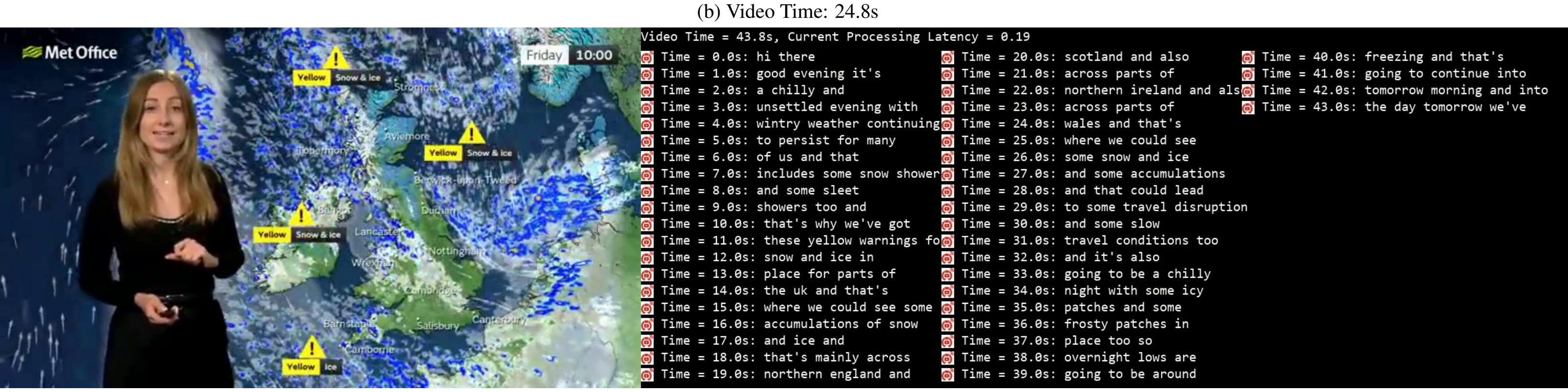
Figure 4. Real-time video commentary demo on unseen YouTube video (`lcZTcfdZ3Ow`). We give the YouTube title "The Planets In Our Solar System" as prompt.



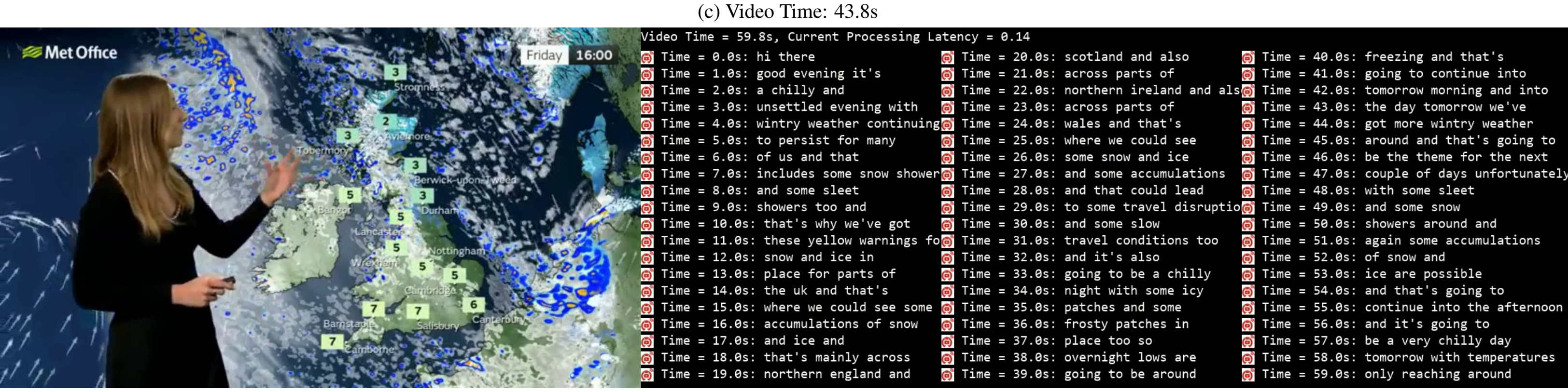
(a) Video Time: 4.3s



(b) Video Time: 24.8s

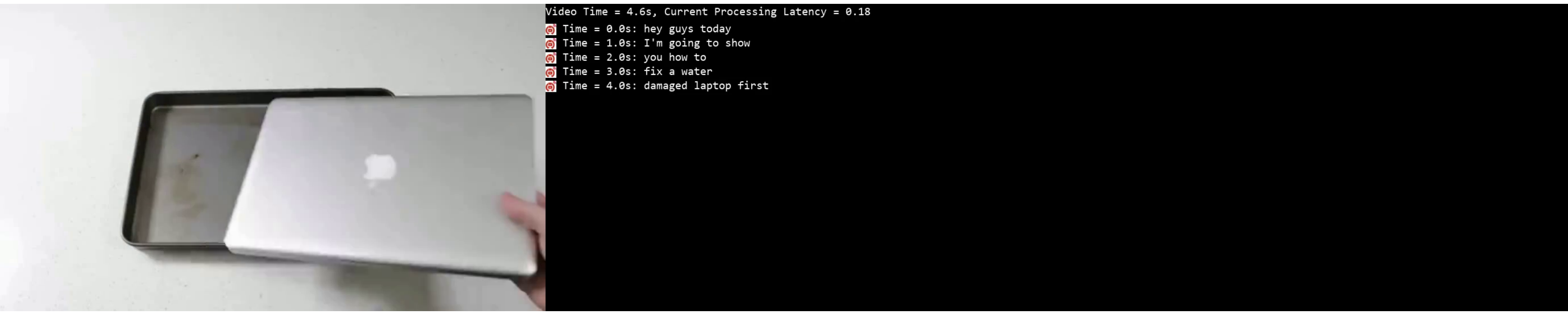


(c) Video Time: 43.8s

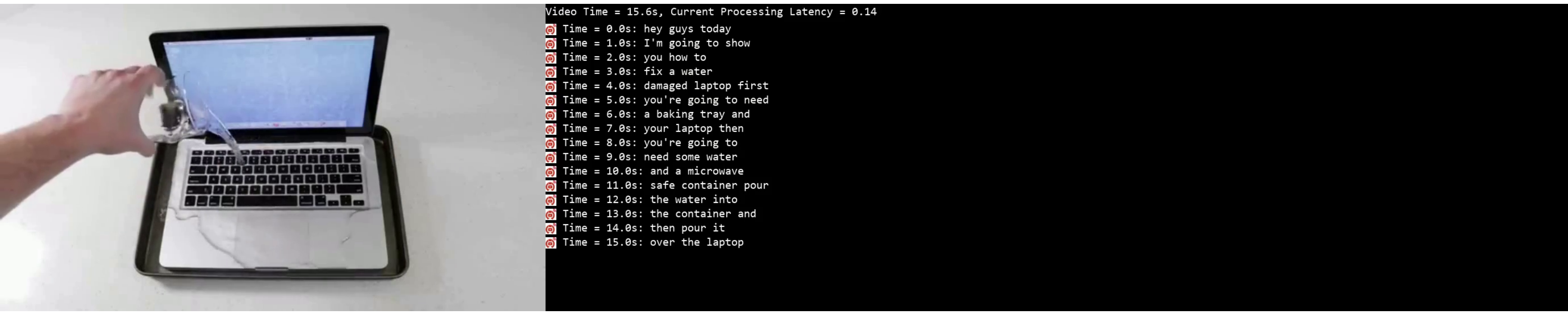


(d) Video Time: 59.8s

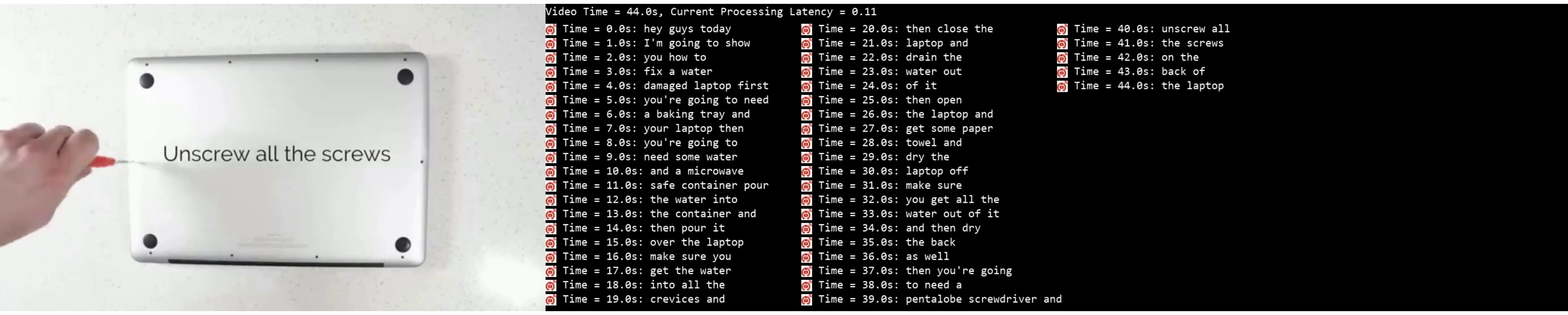
Figure 5. Real-time video commentary demo on unseen YouTube video (`8XajZdrCDsk`). The original YouTube title is "21/11/24 - Wintry weather perservering - Evening Weather Forecast UK – Met Office Weather". We only give "21/11/24 - Wintry weather perservering" as prompt to avoid information leakage.



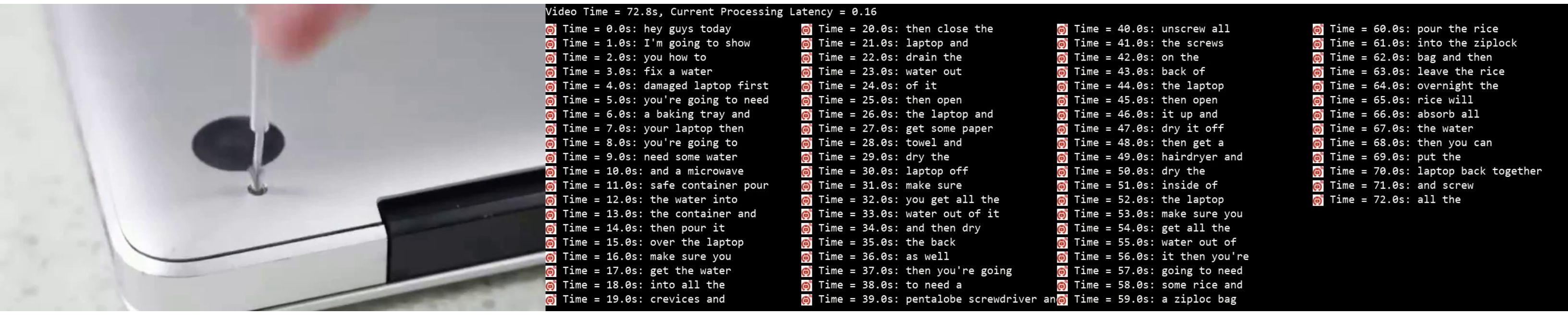
(a) Video Time: 4.6s



(b) Video Time: 15.6s



(c) Video Time: 44.0s



(d) Video Time: 72.8s

Figure 6. Real-time video commentary demo on unseen YouTube video (`115amzVdV44`). We give the YouTube title “How To Fix a Water Damaged Laptop” as the prompt.

References

[1] 2noise. Chattts, 2024. GitHub repository. 1

[2] Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. Gemini: A family of highly capable multimodal models. *arXiv:2312.11805*, 2023. 1, 3

[3] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[4] GPT-4o. Hello gpt-4o, 2024. 1

[5] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[6] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024. 1

[7] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*, 2024. 1

[8] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv:2407.12772*, 2024. 2

[9] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024. 1

[10] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024. 1

[11] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv:2410.02713*, 2024. 1

[12] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 3