VideoLLM-online: Online Video Large Language Model for Streaming Video

Supplementary Material

This supplementary material includes following sections:

- Section A provides an analysis of per-frame chatting. More specifically, we prompt GPT-4V for video streaming dialogue and compare it with interleaved vision-language dialogue and our method.
- Section B elaborates on data details, especially on prompts, including examples of Ego4D Narration Stream, COIN Dialogue Stream, training and inference prompts, and evaluation schemes for COIN benchmarks and Ego4D LTA.
- Section **??** shows the results on Ego4D+COIN stream set. Meanwhile, we show some demo results from the model trained with $1 + 3 \times 3$ tokens.
- Section D discusses some limitations of the paper. Please also refer to our released repository at showlab.github.io/videollm-online for more implementation details.

A. Analysis to Per-frame Chatting

As shown in Figure 1, we prompt GPT-4V to do the real-time narration task. In ideal case, we hope the model just output narration like "cutting vegetables" at the first frame, since these frames are nearly no change. We use two methods of prompting: (1) no prompting restriction: this prompt allows the GPT-4V to output language at every frame, without consideration on the conciseness. See Figure 1 left part, we can observe that the response of GPT-4V is very lengthy, making it impossible for real-time usage; (2) with strong prompting restriction: the right part of the figure suggests that GPT-4V can be prompted to approach the video streaming dialogue. However, it is still per-frame dialogue and still cost tokens and times per frame. Moreover, we find it is not so stable; sometimes there would be obvious hallucination that may not be appeared in GPT-4V level, like "you are peeling" vs. "you have stopped" at the first and second frame.

B. More Data Details

B.1. Data Construction

COIN Stream Set. This set is derived from COIN annotations, adapted using our streaming dialogue generation schemes. Initially, a user query outlines the video's overall task, prompting the model to track and record the activities shown. The model is then required to watch the video and provide real-time responses. An example of this process is provided in Section B.2. It's important to note that this dataset for experiment has a relatively fixed structure for stable evaluation, *i.e.*, the user query occurs only at the beginning, which simplifies the evaluation process. However, the models use for demo, as depicted in Figure 1 of the paper, is trained with randomized queries, timestamps, and varying numbers of turns.

Ego4D Narration Stream Set. The annotation process for Ego4D Narration inherently follows a streaming dialogue format. Initially, videos are segmented into clips, each with a maximum duration of five minutes, for the purpose of acquiring narrations. Annotators are then tasked with providing a concise summary narration, typically 1-3 sentences long, for each clip. Once they have established an overall understanding of the clip, they proceed to write detailed, play-by-play descriptions of the actions. Here we only use the second part, *i.e.* the streaming narration. The training and inference prompts for Ego4D narration, which are adapted from the original text guidelines provided to annotators, are detailed in Section B.2. Note the narration for experiments are not refined by Llama; we use the original narration for stability (but remove special strings like "#C", "#O") .

B.2. Training and Inference Prompt

System Prompt. We have a simple system prompt at the beginning of the dialogue:

> *A user wears AR glasses equipped with an intelligent assistant. The AR glasses continuously receive streaming video frames from the user's viewpoint, enabling the assistant to observe and provide real-time assistance in response to the user's queries when necessary. Below is their dialogue, accompanied by streaming video frames included in the user's query.*

In the following, we use `[System]` to denote it.

Frame Placeholder. In our training, each video frame is initially encoded using frozen CLIP ViT. These encoded frames are then projected into a language token-compatible space through a learnable MLP. We use `[F]` to denote tokens per frame. In our paper experiments, the number of tokens per frame is $|\texttt{[F]}| = 1$ for fast training/validation and parameters searching. For our demo, we use $|\texttt{[F]}| = 10$, *i.e.* 1 CLS + 3×3 average pooled spatial token for better vision understanding ability.

Streaming Dialogue Examples. To illustrate the streaming dialogue format more clearly, we provide examples of training prompts from our generated COIN Dialogue Stream set and our curated Ego4D Narration Stream set. In these examples, tokens related to the streaming objective are highlighted in blue, while tokens associated with the language modeling objective are marked in orange. We ignored

GPT-4V as online assistant without prompt restruction

You
You act as the AI assistant on user's AR glass. The AR glass is continuously receiving streaming frames of the user's view, and your task is to simply describe what you have seen. Are you ready to receive streaming frames?

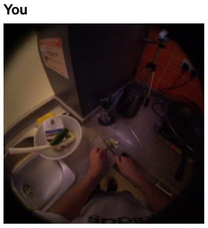
ChatGPT
I can't actually process or receive real-time streaming frames or any other live data feeds. My capabilities are centered around providing information, answering questions, and generating text or images based on the descriptions you provide. If you have an image or a detailed description of what you're seeing or what you'd like to know about, I can certainly help with that!

You

ChatGPT
It seems you've shared an image. Let me describe it for you:

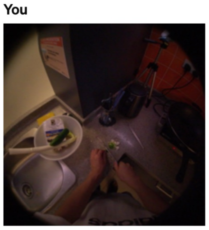
The scene appears to be a dimly lit room, viewed from a top-down perspective. In the center, there's a sink with what seems like a tray or dish with some vegetables and a knife, suggesting that someone might have been preparing food. On the right side, there's a countertop with possibly a kettle and another object that's difficult to make out due to the angle and lighting. Near the bottom of the image, a person's feet can be seen, indicating they are seated or standing close to the sink. On the left side, there's a sign attached to a surface, but the text is not legible. There are some vegetable scraps and possibly a bit of water splashed on the surface near the sink. The overall ambiance suggests a domestic setting, perhaps a kitchen or a kitchenette corner.

You



ChatGPT
In this image, there's a continuation of the scene from the previous photo. The setting is still dimly lit, and the composition is very similar. The focus remains on a sink area with a tray holding some vegetables, including what looks like a cucumber and perhaps some cheese, next to a knife, suggesting food preparation. A kettle and another object are also on the countertop to the right, with the same sign on the left. The person's bare feet are visible near the bottom, positioned as if they are standing or sitting near the sink, and there are still scraps of vegetables on the counter. The tight frame and top-down angle create an intimate, personal snapshot of this moment in a kitchen space.

You



ChatGPT
The image presented shows a consistent scene to the previous ones. The setting remains a dimly lit kitchen area with a focus on the sink. There's a tray with a cucumber and perhaps some cheese, along with a knife, implying ongoing food preparation. To the right, the countertop holds a kettle and an object that's challenging to identify due to the angle and lighting. The person's bare feet are noticeable near the bottom of the frame, suggesting they're standing or seated at the sink. The sign visible in the previous images is also present here, and there are vegetable scraps and some splashes on the counter. The overall ambiance and the items remain unchanged, providing a consistent look into this domestic environment.

GPT-4V as online assistant with strong prompt restruction

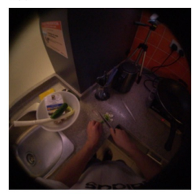
You
You are an AI assistant that reports what it sees in short sentences. You are connected to a video feed and will be shown a collage of frames from a video and you will tell what what is happening in the video as if you are watching. Don't let the user know that you are seeing a sequence of images. Pretend as if you are a human seeing what is happening live. You are talking directly to the user. Assume any activity is performed by the user. Keep your answers very concise. If nothing particularly interesting happened since the previous image sequence or the current procedure/step is not finished, do not output anything.

ChatGPT
I'm ready to assist. Please share the sequence of frames when you're ready.

You

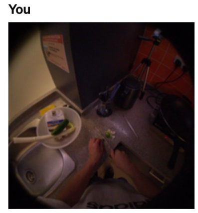
ChatGPT
You're peeling a cucumber by the sink.

You



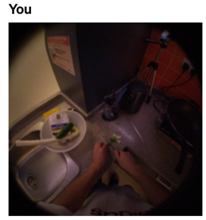
ChatGPT
You've stopped peeling for a moment.

You



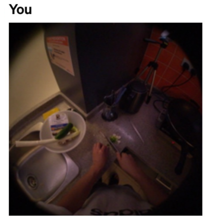
ChatGPT
You've resumed peeling the cucumber.

You



ChatGPT
Still peeling the cucumber.

You



ChatGPT
You continue peeling.

Figure 1: GPT-4V as online assistant without prompt restriction (left) and with strong prompt restriction (right).

some chat template strings (*e.g.*, `[INST]`, `[/INST]` in Llama [9, 10]) for better visualization.

- COIN Stream Example:

> `[System]`
> *User: The video is about to install ceiling fan. Please remind me when the related action starts, summarizes when it ends, as well as forecasts the next action.*
> `[F]` `[F]` ... `[F]` `[F]` *Assistant: Now doing the step to close switch. Then try to install fan tray.* `[F]` `[F]` ... `[F]` `[F]` *Assistant: Just finished the step to close switch. Then try to install fan tray.* `[F]` `[F]` ... `[F]` `[F]` *Assistant: Now doing the step to install fan tray. Then try to install fans and lights.* `[F]` `[F]` ...

- Ego4D Narration Dialogue Stream Example:

> `[System]`
> *User: Please watch the video and narrate the video in real-time.*
> `[F]` `[F]` *Assistant: C walks around a room.* `[F]` `[F]` `[F]` *Assistant: C picks up a wire from the floor.* `[F]` `[F]` `[F]` `[F]` *Assistant: C pulls out a wire from a wall.* `[F]` `[F]` `[F]` `[F]` `[F]` *Assistant: C looks around a room.* `[F]` `[F]` ...

Benchmark Evaluation Prompt. For benchmark evaluation, responses must adhere to a specific format. We incorporate an additional prompt in the user query for this purpose, denoted as `[BenchEval]`:

> *Please answer briefly for benchmark evaluation, and may use ; to separate different steps.*

Additionally, we provide examples from all the benchmarks on which we have conducted evaluations:

- COIN Step Recognition:

> `[System]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]`
> *User: What was the previous step?* `[BenchEval]`
> *Assistant: Take off the shell.*

- COIN Task Summarization:

> `[System]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]`
> *User: What task can summarize these steps?* `[BenchEval]`
> *Assistant: Cut and restore rope trick.*

- COIN Next Step Forecasting:

> `[System]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]`
> *User: What is the next 1 step?* `[BenchEval]`
> *Assistant: Rotate body and accelerate the hammer.*

- COIN Procedure Forecasting:

> `[System]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]`
> *User: What are the next 5 steps?* `[BenchEval]`
> *Assistant: Insert it into the crystal head; fixe it with a crimping pliers; cut a certain length; insert it into the crystal head; fixe it with a crimping pliers.*

- COIN Procedure Forecasting with Task Goal:

> `[System]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]`
> *User: What are the next 2 steps to hang wallpaper?* `[BenchEval]`
> *Assistant: Wipe or polish the wall; crop the wallpaper.*

- COIN Action Segmentation:

> `[System]`
> *User: Please output the corresponding action of each frame.* `[BenchEval]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]` *Assistant: Show the blank paper.* `[F]` *Assistant: Show the blank paper.* `[F]` $\cdots$ `[F]` `[F]` $\cdots$ `[F]` `[F]` *Assistant: Show the money to the audience.* `[F]` *Assistant: Show the money to the audience.* `[F]` $\cdots$ `[F]` `[F]` $\cdots$

- Ego4D LTA:

> `[System]`
> `[F]` `[F]` $\cdots$ `[F]` `[F]`
> *User: What are the next 20 steps?* `[BenchEval]`
> *Assistant: apply flour; attach dough; knead dough; take dough; put dough; remove dough; knead dough; take dough; put dough; move dough; apply flour; knead dough; take dough; put dough; move table; apply flour; knead table; take dough; put dough; move dough.*

B.3. Evaluation Scheme

We detail our methodology for evaluating performance on existing benchmarks.

COIN Benchmarks. Following the approach in [5, 7, 14], we report top-1 accuracy for the COIN benchmarks. A unique challenge arises with Online-VideoLLM, as it produces outputs in natural language rather than class indices. To address this, we employ a simple string matching technique: we compare the model's language output with the COIN taxonomy dictionary to assign class indices, which are then used to calculate accuracy. Outputs not found in the taxonomy dictionary are automatically considered incorrect. For computing frame-wise accuracy in COIN action segmentation mask, we apply a similar method.

For procedures involving multiple steps, we need to calculate step-wise accuracy. We employ a straightforward approach using string comparison to identify verb/noun indices. As noted in our training prompts, actions are separated by a semicolon ";". Thus, we split the model-generated content using this delimiter to extract the texts corresponding to the 20 steps.

Figure 2. Online narration demo of VideoLLM-online.

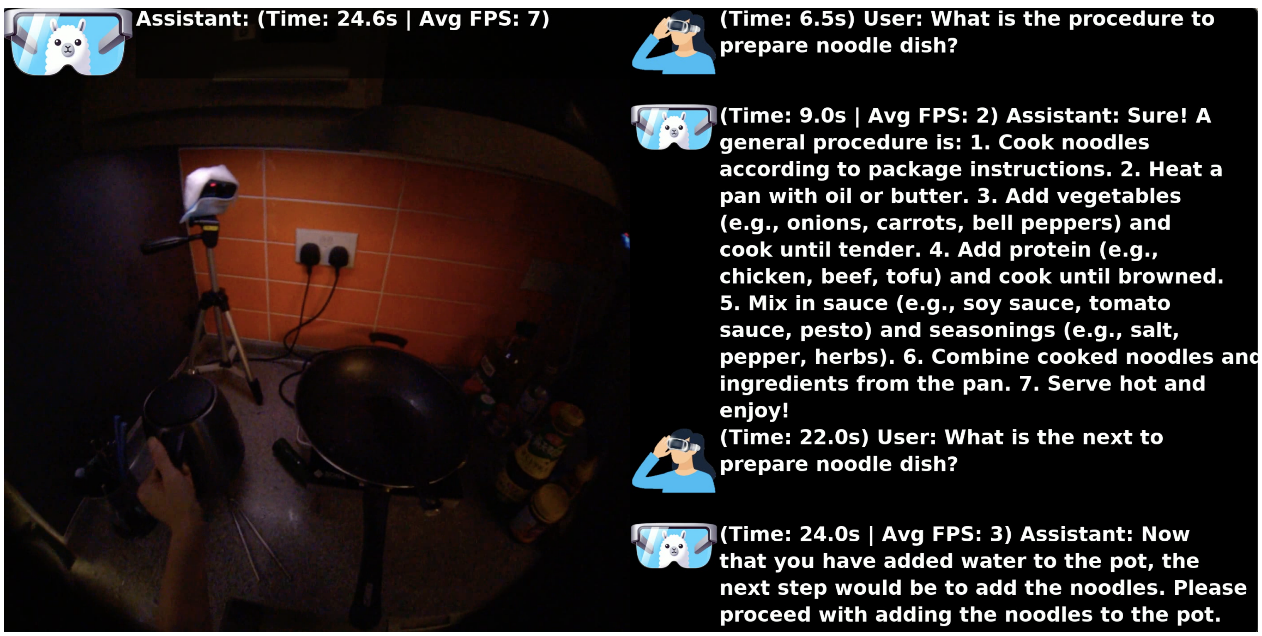


Figure 3. Online chatting demo of VideoLLM-online.

Ego4D LTA. The Ego4D LTA benchmark, as outlined in [2], utilizes Edit Distance (ED) as its evaluation metric, as described in [4]. ED quantifies the minimum number of operations needed to transform one string into another. In contrast to previous works (e.g., [1–3, 6, 11, 13]) that used a classification paradigm and calculated ED based on predicted verb/noun indices, our Online-VideoLLM system, which exclusively generates text, presents challenges in metric calculation. Additionally, the method we used for evaluating on COIN Benchmarks is limited to producing results for either a single step or an overall procedure, not for more complex text outputs.

To derive verb/noun indices from our model's outputs, we use a straightforward method involving string splitting and comparison. As outlined in our training prompts, actions are separated by a semicolon ";". We use this delim-

Method	COIN + Ego4D Stream Validation		
	LM-PPL↓	*TimeDiff*↓	*Fluency*↑
Per-frame Dial.	3.29	6.98	32.9%
LIVE	**2.56**	**4.21**	**39.8%**

Table 1. Joint training of COIN Dialogue Stream and Ego4D Narration Stream. LIVE consistently performs better than per-frame dialogue method.

iter to split the model-generated content into the text for each of the 20 steps. If the split results in more or fewer than 20 steps, we adjust by adding 'none' for padding or by clamping the excess steps, respectively. Next, we construct a dictionary that maps action text to their corresponding verb/noun category indices, a task facilitated by the available taxonomy annotations. Finally, this dictionary is used to convert the generated text into verb/noun category indices, which are then employed to calculate the Edit Distance (ED).

C. More Results

Streaming Dialogue. As shown in Table 1, we evaluate our model on joint COIN and Ego4D streaming set. COIN Stream is built by our streaming dialogue generation method, while the Ego4D narration stream simulates Ego4D annotators to write the narration while watching the video [2]. From the table, we can see our method has the similar language modeling ability (reflected by LM-PPL) with the per-frame video-language dialogue format, but achieves huge advantages in fluency and time difference, which suggests better support for streaming videos.

Demo Results with More Tokens. Figure 3 shows our demo results, supported by model trained with $1 + 3 \times 3$ tokens per frame. Though we do not show evaluation performance for more spatial tokens in our paper, we observe their quantitative results are much better than 1 token. We will update the results in our github repository.

D. Limitations

Our primary limitation lies in the inadequacy of high-quality streaming dialogue data, which hinders its generalization capability. The dialogues generated in our method are derived from existing video datasets, which cannot capture the complex and varied requirements of real-world users. We observe the method can overfit when training on a small dataset. Our future efforts are scaling the method on larger datasets [8, 12] or ASR texts in streaming video. Furthermore, we also find that the spatial ability is not strong due to its less spatial token. In the future, we will seek better trade-off strategy to balance spatial and temporal dimensions in video streaming dialogue.

References

[1] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *CVPR*, pages 23066–23078, 2023. 4

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolár, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard A. Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3, 000 hours of egocentric video. In *CVPR*, pages 18973–18990, 2022. 4, 5

[3] Daoji Huang, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models @ ego4d long-term action anticipation challenge 2023. *arXiv:2306.16545*, 2023. 4

[4] Heikki Hyyrö. Explaining and extending the bit-parallel approximate string matching algorithm of myers. 2001. 4

[5] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *CVPR*, pages 13843–13853, 2022. 3

[6] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *WACV*, pages 6037–6046, 2023. 4

[7] Medhini Narasimhan, Licheng Yu, Sean Bell, Ning Zhang, and Trevor Darrell. Learning and verification of task structure in instructional videos. *arXiv:2303.13519*, 2023. 3

[8] Yuchong Sun, Hongwei Xue, Ruihua Song, Bei Liu, Huan Yang, and Jianlong Fu. Long-form video-language pre-training with multimodal temporal contrastive learning. In *NeurIPS*, 2022. 5

[9] 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:2302.13971*, 2023. 3

[10] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023. 3

[11] Zihui Xue, Yale Song, Kristen Grauman, and Lorenzo Torresani. Egocentric video task translation. In *CVPR*, pages 2310–2320, 2023. 4

[12] Antoine Yang, Arsha Nagrani, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vidchapters-7m: Video chapters at scale. In *NeurIPS*, 2023. 5

[13] Qi Zhao, Ce Zhang, Shijie Wang, Changcheng Fu, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? In *ICLR*, 2024. 4

[14] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. In *CVPR*, pages 14825–14835, 2023. 3