

# What to Say and When to Say it: A Video-Language Model and Benchmark for Situated Interactions

Sunny Panchal<sup>\*1</sup>, Apratim Bhattacharyya<sup>\*1</sup>, Guillaume Berger<sup>1</sup>, Antoine Mercier<sup>1</sup>, Cornelius Böhm<sup>2</sup>, Florian Dietrichkeit, Xuanlin Li<sup>3</sup>, Reza Pourreza<sup>1</sup>, Pulkit Madan<sup>1</sup>, Mingyu Lee<sup>1</sup>, Mark Todorovich<sup>1</sup>, Ingo Bax<sup>1</sup>, Roland Memisevic<sup>1</sup>

<sup>1</sup>Qualcomm AI Research<sup>†</sup>    <sup>2</sup>Aignostics GmbH<sup>‡</sup>    <sup>3</sup>UC San Diego<sup>§</sup>

## Abstract

*Large language models have flourished as general purpose dialogue-based assistants, however, they remain limited to turn-based interactions about offline documents or images. Open-ended, asynchronous interaction with situated agents is an open challenge. Recent advances in multi-modal models take a step towards giving these assistants eyes and ears, but they are limited to capturing momentary snapshots of reality in a VQA-style dialogue. In this work, we take a step towards situated vision-language models that can process a live video stream and can dynamically interact with users. To this end, we present the FIT-COACH benchmark and dataset. It is based in the controlled but challenging fitness coaching domain and is aimed at the development of interactive multi-modal vision-language models. Alongside this, we present a novel interactive video-language model, COACH-LLAMA, that can respond inter-actively to events in the visual input stream.*

## 1. Introduction

Tasks that combine vision and language [3, 7, 9] have had a profound impact on computer vision. Despite recent advances, vision-language models still fall short of human capabilities in terms of visual understanding. A key distinction is their lack of temporal experience. Most models process an image or video with an instruction, producing a single output. Even visual dialogue, the most complex vision-language task, remains “turn-based” and controlled externally through prompting, unlike fluid human conversation (e.g., by entering a new-line character, Fig. 1 left).

This is in stark contrast to many real-world dialogues, in which participants decide what to say and when to speak,

based in equal parts on what has been said in the past and on the visual scene unfolding in real-time in front of their eyes. Such real-world dialogues are commonly referred to as “situated” [2, 4, 5]. Situated interactions are far from a solved problem in AI, although progress in this area would enable many novel real-world applications, including coaching and instruction dialogues, as well as collaborative tasks.

In this work, we introduce a real-world benchmark as a highly controlled, albeit in-the-wild, test-bed for studying situated vision-language generation. We use live fitness coaching as an interaction scenario, where an AI model interacts with a user through language, while having real-time access to visual information through a camera. This real-world scenario has several benefits that make it an ideal test-bed for studying situated dialogue: First, it is highly controlled and game-like, in that users are expected to follow a specific series of movements while models are expected to coach them. Second, despite its controlled structure, it is a highly challenging unsolved problem for current AI systems. Third, coaching is a real-world computer vision application, solutions to which could provide real and tangible benefits to users.

Overall, our main contributions are: 1) We propose the first large scale benchmark and dataset, FIT-COACH, aimed at the development of video language models for situated interactions. Our FIT-COACH dataset and benchmark contains 470+ hours of videos. It includes short-clip videos (~5 sec in length) annotated with 1.7M+ question-answer pairs, and long-range videos (>3 minutes in length) annotated with live feedbacks; 2) We propose a novel video LM, COACH-LLAMA, for situated interactions. As shown in Fig. 1, unlike current state-of-the-art video LMs, it is not constrained to turn-based interactions, but instead is able to determine on-the-fly when and what to say to the user; 3) We show that our proposed COACH-LLAMA model outperforms state-of-the-art video language models on the FIT-COACH benchmark.

<sup>\*</sup>Authors contributed equally

<sup>†</sup>Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.

<sup>‡</sup>Work performed at TwentyBN GmbH

<sup>§</sup>Work performed at Qualcomm AI Research

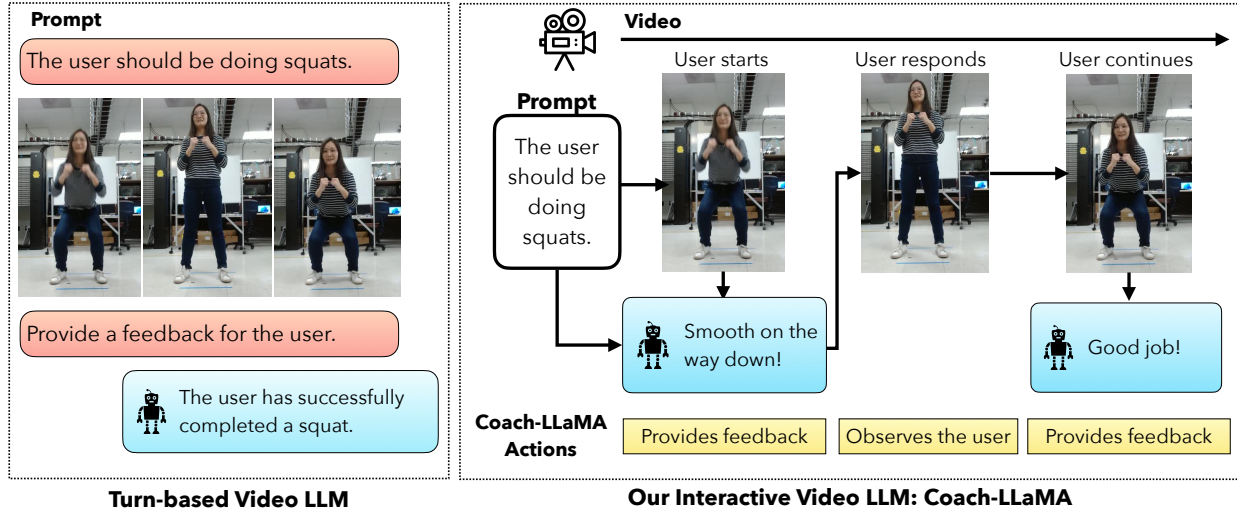


Figure 1. Current state-of-the-art video-language models are largely “turn-based” and respond only when specifically prompted. In contrast, our COACH-LLAMA model can respond interactively to a real-world video stream input.

## 2. Fitness Interactive Coaching Benchmark

To build an end-to-end video-language model (LM) for situated interaction, a key challenge lies in providing appropriate responses (“what to say”) at the correct time (“when to say it”). The Fitness Interactive Coaching (FIT-COACH) dataset and benchmark addresses this. It offers 470+ hours of annotated short videos for low-level visual concept formation, as well as 10+ hours of long-range videos with live fitness coaching for timing-based guidance. The FIT-COACH benchmark focuses on situated interaction within a subset of these long-range videos. Full dataset details are in the appendix.

The FIT-COACH benchmark contains videos of participants performing a structured workout and timestamped feedbacks. These live feedbacks serve to point out any mistakes and to provide encouragement to the user; examples of such feedbacks can be seen in Fig. 2. In the workout session shown in Fig. 2 (top), while performing squats the participant receives feedback to ensure they are going deep enough. Later in the session, a positive comment is made on the participant’s smooth movement, encouraging them to continue doing so. Next, while performing jumping jacks, the user receives positive reinforcement for good performance and encouragement to continue performing more repetitions of the exercise. Finally, while doing the high knees exercise, the participant receives feedback to increase their speed followed by positive reinforcement with an encouraging comment when they do. These examples highlight the highly interactive workout coaching sessions in our FIT-COACH benchmark and showcase the tight coupling between participant actions and timely feedback.

**Annotations:** Annotations include feedbacks which are generated semi automatically using a rule-based approach.

The rule-based approach detects events such as common mistakes made by participants as well as their form and speed. When an event is detected it triggers the generation of a feedback (more details in the supplemental). If multiple feedbacks are triggered at the same time, a single one is randomly selected. The feedbacks are post-processed manually to improve diversity, verify correctness, and ensure long-range consistency. Careful attention is paid to ensure edits don’t corrupt causal interactions.

**Statistics:** In total, the FIT-COACH benchmark consists of  $\sim 4.5$  hours of recorded workout sessions. Each session is  $\sim 3.5$  minutes long and consists of 5 to 6 randomly selected exercises out of 23 possible exercises. The full list of exercises is provided in the supplemental. There are a total of 7 unique participants with a cumulative recording length of  $\sim 20$  minutes to  $\sim 1.5$  hours each.

## 3. COACH-LLAMA

Here, we describe our Coach-LLaMA model, capable of interactive feedback generation on a continuous vision stream, such as exercise coaching. The model consists of a 3D CNN-based vision backbone (based on [1]) for processing the vision stream, a pre-trained LLaMA-2-7B LM backbone to generate interactive feedbacks, and a cross-attention-based adapter fusing the two. In order to respond interactively to events observed in the vision stream, we expand the vocabulary of the backbone LM to include special action tokens. We provide details of the vision backbone and adapter in the appendix.

### 3.1. Action Tokens

Our COACH-LLAMA model uses two special action tokens `<next>` and `<feedback>` to enable interactive feed-

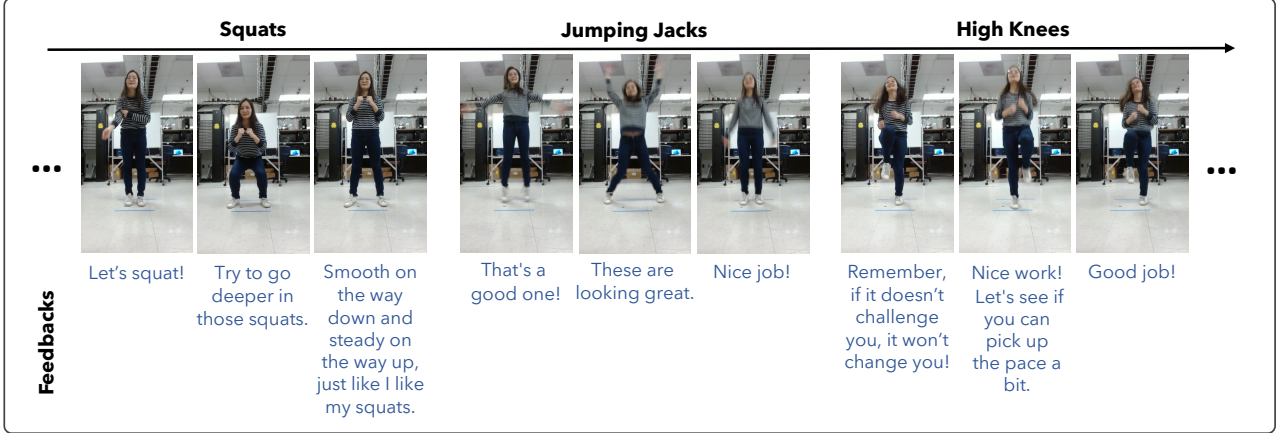


Figure 2. Examples of the long-range interactive videos from our FIT-COACH benchmark. The live feedbacks provided to the participants are shown below each frame.

backs. The `<next>` token allows our model to request the next video frame as input from the visual stream. When the model decides to provide feedback, it outputs the `<feedback>` token. This is illustrated in Fig. 4, where our COACH-LLAMA model guides a user through a squats exercise. It observes the user for a few repetitions by requesting frames from the visual stream using the `<next>` token. Then, at the time-step it decides to provide feedback, it outputs the `<feedback>` token. This is followed by the feedback: “Smooth on the way down . . .”. After the model is finished providing feedback, it requests the next video frame using the `<next>` token.

Note that, depending on the amount of time required to produce the feedback tokens, a few frames might not be observed by our COACH-LLAMA model while the feedback is being generated. Empirically, we do not observe any degradation in performance on commercial hardware, as the number of frames dropped are few and the user typically requires some time to react meaningfully to any provided feedback. Thus, overall the `<next>` and `<feedback>` special action tokens enable the model to consistently provide appropriate feedback in a sufficiently timely manner.

## 4. Experiments

Next, we evaluate our COACH-LLAMA model on the interactive FIT-COACH benchmark.

### 4.1. Evaluation Metrics

We use the following metrics that emphasize both the fluency and temporal accuracy.

**Temporal-F-Score.** We measure whether predicted responses occur at the correct time-step using the temporal F-score. Predicted responses within a coaching session are classified as true or false positives by temporally aligning them with ground truth responses. We match each ground

Method	T-F-Score $\uparrow$	T-BERT $\uparrow$	T-Rouge-L $\uparrow$
Socratic-Llama-2-7B $^\dagger$	0.57	0.441	0.041
InstructBLIP $^\dagger$ [6]	0.57	0.437	0.027
Video-LLaMA $^\dagger$ [15]	0.57	0.436	0.029
Video-ChatGPT $^\dagger$ [12]	0.57	0.439	0.033
COACH-LLAMA (Ours)	<b>0.64</b>	<b>0.512</b>	<b>0.115</b>
COACH-LLAMA (w/o 3D CNN)	0.56	0.361	0.058
COACH-LLAMA (w/o Action Tokens)	0.57	0.428	0.088
COACH-LLAMA (w/o Pre-training)	0.54	0.383	0.073

Table 1. Evaluation on the FIT-COACH benchmark. ( $^\dagger$  indicates off-the-shelf models.)

truth response to the closest predicted response within a  $\pm 2$  second window, maintaining their temporal order. Predicted responses with a ground truth match are true positives, otherwise false positives. Ground truth responses without a match are false negatives. This allows us to calculate temporal precision, recall, and thus the temporal F-Score.

**Temporal-BERT/Rouge-L Scores.** The temporal-BERT and temporal Rouge-L scores are calculated by first aligning ground truth and predicted feedbacks, as done for the temporal-F-score. For each aligned pair, we compute standard BERT/Rouge-L scores [11, 16]. Unaligned ground truth feedbacks lead to a score of zero. The final score is the average across all aligned pairs and these (zero-score) unaligned ground truth feedbacks. Unlike standard versions, temporal-BERT and temporal-Rouge-L factors in both the temporal alignment and the fluency.

### 4.2. Evaluation on the FIT-COACH Benchmark

We begin by evaluating our COACH-LLAMA model on the FIT-COACH benchmark in Tab. 1. We compare to state-of-the-art (open-source) multi-modal language models InstructBLIP [6], Video-LLaMA [15] and Video-ChatGPT [12]. As these models are not interactive, we prompt them to provide feedbacks at regular intervals (Fig. 1) in a turn-

based fashion. We found that a regular interval of 6 seconds leads to the best overall performance of these turn-based models. We also consider a “text-only” socratic model [14], where we prompt the language-only LLaMA-2-7B LM with descriptions of user activity at 1-second intervals for the previous 6 seconds to generate a response. The textual description of user activity is based on the aforementioned feedback event triggers.

We also consider the following ablations of our COACH-LLAMA model: 1) Instead of our 3D CNN, we use a CLIP based encoder similar to Video-ChatGPT [12] (w/o 3D CNN); 2) A non-interactive turn-based version of our COACH-LLAMA model without the <next> and <feedback> action tokens (w/o Action Tokens); 3) Training only on the long-range videos from our fitness feedbacks dataset, without pre-training on the short-clip videos of our fitness questions and feedbacks dataset (w/o Pre-training).

The results in Tab. 1 demonstrate that our COACH-LLAMA model significantly surpasses the performance of current state-of-the-art models like InstructBLIP, Video-LLaMA, and Video-ChatGPT in temporal F-score. Our model’s superior temporal F-score indicates its ability to provide feedback at the appropriate time-steps, unlike turn-based models. This is further reinforced by the significant drop in performance of the COACH-LLAMA (w/o Action Tokens) ablation.

We also expose a key limitation of the CLIP-based video encoders used in existing models – their inability to capture the fine-grained motion details crucial for accurate fitness coaching feedback. This limitation is further highlighted by the poor performance of the COACH-LLAMA (w/o 3D CNN) baseline.

Another essential component of our COACH-LLAMA model is the pre-training on the fitness questions and (short-clip) feedbacks datasets. This pre-training provides the domain knowledge necessary for detailed feedback generation. Qualitative examples of interactive feedbacks generated by the model are shown in Fig. 5 (top): the user starts performing jumping jacks but stops partway through. In this case, the model correctly encourages the user to continue. It also successfully detects that the user is only using their arms and later acknowledges their correction.

The feedbacks that can be generated from existing off-the-shelf models, such as InstructBLIP, Socratic-LLaMA-2-7B, Video-LLaMA, and Video-ChatGPT, by comparison are more generic and lacking in specifics (see supplemental) due to their lack of domain knowledge, as discussed in Appendix B.3. This is further highlighted in Tab. 4, where due to the feedbacks lacking specific details, we see significantly lower temporal-ROUGE-L scores. This underscores the utility of our FIT-COACH dataset in the development of interactive video-language models for fitness coaching.

## 5. Conclusion

We propose FIT-COACH, a novel interactive visual coaching benchmark and dataset as a test-bed for real-time, real-world situated interaction. Alongside this, we present COACH-LLAMA, a novel vision-language model that can learn not only what, but also when, to reply to a user based on a real-world video stream. We consider this work as a starting point for further research into end-to-end training of domain-specific interactive vision models, and hope that our data and baselines will encourage further work.

## References

- [1] Sense Core. <https://github.com/quic/sense>. [Online; accessed 17-Nov-2023]. 2, 5
- [2] Prithviraj Ammanabrolu, Renee Jia, and Mark O Riedl. Situated dialogue learning through procedural environment generation. In *ACL*, 2022. 1
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1
- [4] Dan Bohus, Sean Andrist, Ashley Feniello, Nick Saw, Mihai Jalobeanu, Pat Sweeney, Anne Loomis Thompson, and Eric Horvitz. Platform for situated intelligence. Technical report, Microsoft, 2021. 1
- [5] Rodney A. Brooks. Intelligence without reason. In *IJCAI*, 1991. 1
- [6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023. 3
- [7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017. 1
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 7
- [9] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In *ECCV*, 2010. 1
- [10] 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. In *ICLR*, 2022. 7
- [11] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, 2004. 3
- [12] Muhammad Maaz, Hanoona Abdul Rasheed, Salman H. Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *CoRR*, abs/2306.05424, 2023. 3, 4, 9
- [13] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 5
- [14] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aweek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language. In *ICLR*, 2023. 4
- [15] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP - System Demonstrations*, 2023. 3, 9
- [16] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020. 3



## Appendix

### A. FIT-COACH Dataset

Here we describe the FIT-COACH dataset used for training our Coach-LLaMA model as described in Sec. 3. The FIT-COACH dataset consists of two annotation subsets: the fitness questions dataset and the fitness feedbacks dataset. The fitness questions dataset is designed to instill domain understanding for fitness coaching whereas the fitness feedbacks dataset is designed for providing effective second-person perspective feedback during live coaching sessions.

**Fitness questions dataset.** This is a large video dataset of fitness (148 exercises) and general (49 types) activities comprised of over 300k short-clip videos, totaling 470+ hours, crowd-sourced from over 1900 unique participants in the wild. Participants were provided detailed instructions and an accompanying reference video to perform the exercises and their pre-determined variations. Variations include varied pacing, performing common mistakes, and modified form, as determined through consultation with expert fitness instructors; approximately 10 variations were collected per exercise. Video lengths are in the 2 – 10 seconds range. There are approximately 3500 videos per exercise on average and a total of 1800+ fine-grained classes capturing the exercise variations and general activities. The full list of classes is provided in the supplemental.

Based on the variations collected, we create corresponding question-answer pairs for each video. The questions can be broadly divided into two types: high-level and fine-grained, as shown in Fig. 3. The high-level questions are directed at the overall exercise type and performance of the participant, *e.g.*, in Fig. 3, high-level questions include “What exercise is the user doing?”, “Is the user doing this right?”. Fine-grained questions are designed to teach fine-grained details of exercises performed by the participant, *e.g.*, “Is the user leaning forward?”, or “Is the user not jumping?”.

Overall statistics for the dataset, including a breakdown of the provided splits, can be found in Tab. 2. The distribution of video clip lengths, full list of classes, and class labels are provided in the supplemental.

**Fitness feedbacks dataset.** This dataset is designed to enable effective second-person perspective feedbacks that are sensible in a live fitness coaching interaction. It contains annotations for both the large short-clip videos collection (as discussed in the fitness questions dataset section), and the long-range videos collection (a subset of which forms the FIT-COACH benchmark) used for training our COACH-LLaMA model.

In total, we collect an additional  $\sim 9$  hours of fitness coaching sessions following the same methodology as the FIT-COACH benchmark. The sessions are approximately

$\sim 3.5$  minutes long and consist of 5 to 6 exercises presented as a structured workout. These sessions were recorded with 21 participants, distinct from the benchmark.

In addition to the long-range videos, the fitness feedbacks dataset includes semi-automatically generated feedback annotations on a subset of the short-clip videos collection. For example, in Fig. 3 (bottom), the feedbacks focus on improving the form of the participant, specifically encouraging them to go deeper on the squats and to punch with both arms. We provide several alternative phrasings for each feedback for linguistic diversity.

### B. COACH-LLaMA

#### B.1. Vision Backbone and Adapter

**Vision backbone.** The vision backbone (Fig. 4) is designed to robustly recognize motion cues, crucial for fitness coaching. This is in contrast to current state-of-the-art video-language models which typically use CLIP/ViT based vision encoders that capture scene content rather than motion. In detail, our vision backbone architecture is based on a publicly available 3D CNN [1] which was shown to be able to recognize a wide range of behavior patterns, including simple exercises. It consists of a mix of 2D and 3D convolutional layers, ensuring that the model can pick up on both motion and content information of individual frames for predictions—both of which are relevant to provide appropriate feedbacks. Specifically, it is based on a modified EfficientNet-Lite4 model [13] where a subset of residual blocks are temporally inflated, turning them into causal 3D convolutional blocks. The input spatial resolution is  $256 \times 256$ , ensuring relevant visual information is preserved. Furthermore, the model is a streaming architecture, due to the use of causal convolutions, making it well-suited in the context of live feedback generation for fitness coaching. It is designed to process video at 16fps in real-time.

**Adapter.** The adapter layer concatenates positional embeddings to the features from the penultimate vision backbone layer to preserve spatial information. As the features from the vision backbone are high-dimensional, we employ a projection and downsampling layer that transforms the vision backbone features to match the embedding space dimensionality of the LM backbone, allowing the use of cross attention layers to map visual information to the language backbone. Features from the vision backbone are fused into the LM backbone through cross attention at several layers. The fusion happens after every `<next>` action token predicted by our COACH-LLaMA model (Fig. 4), as described in the following section. Our experiments indicate the full model is capable of running  $2.3\times$  real-time on a single consumer-grade GPU.

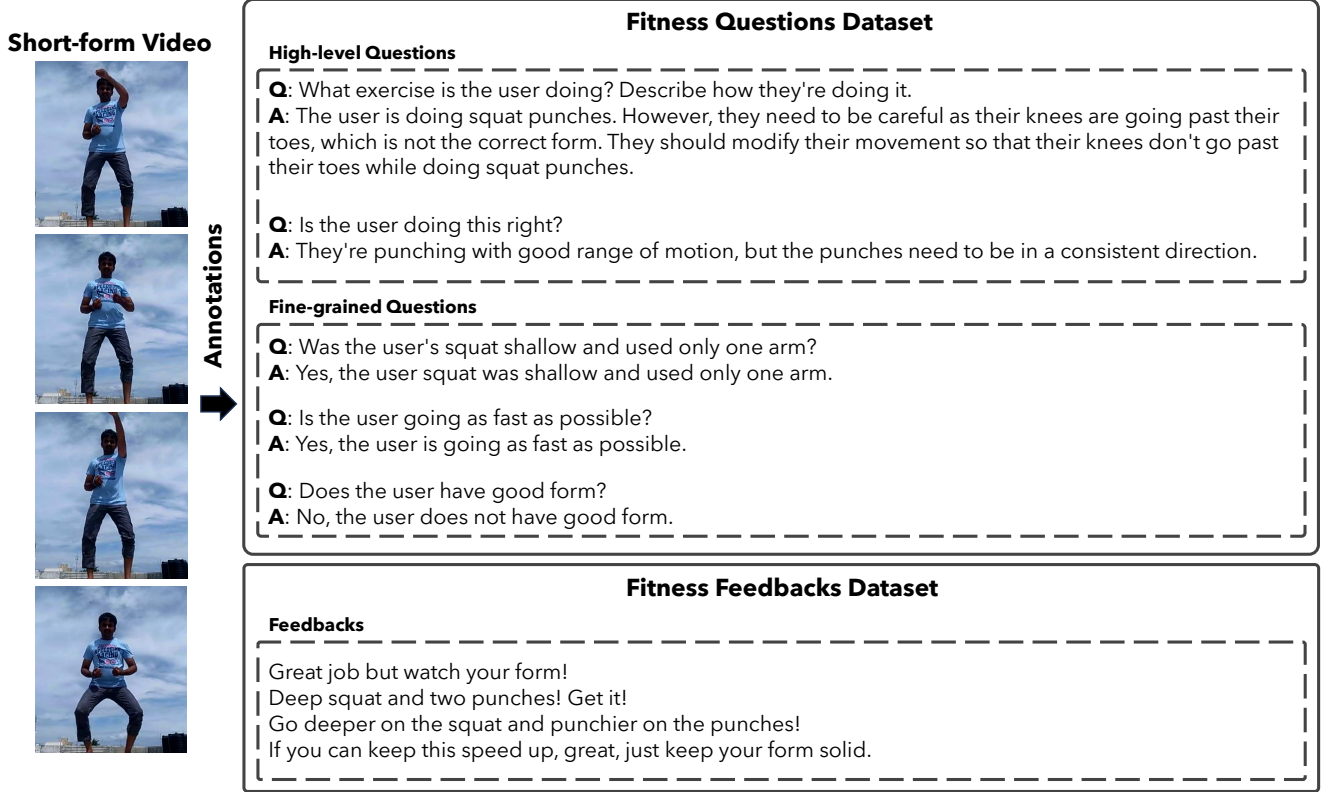


Figure 3. Example annotations available on the short video clips from the FIT-COACH dataset. Annotations include question/answer pairs from our fitness questions dataset and feedbacks from our fitness feedbacks datasets. The fitness feedbacks dataset also contains annotations on long-range videos similar to Fig. 2

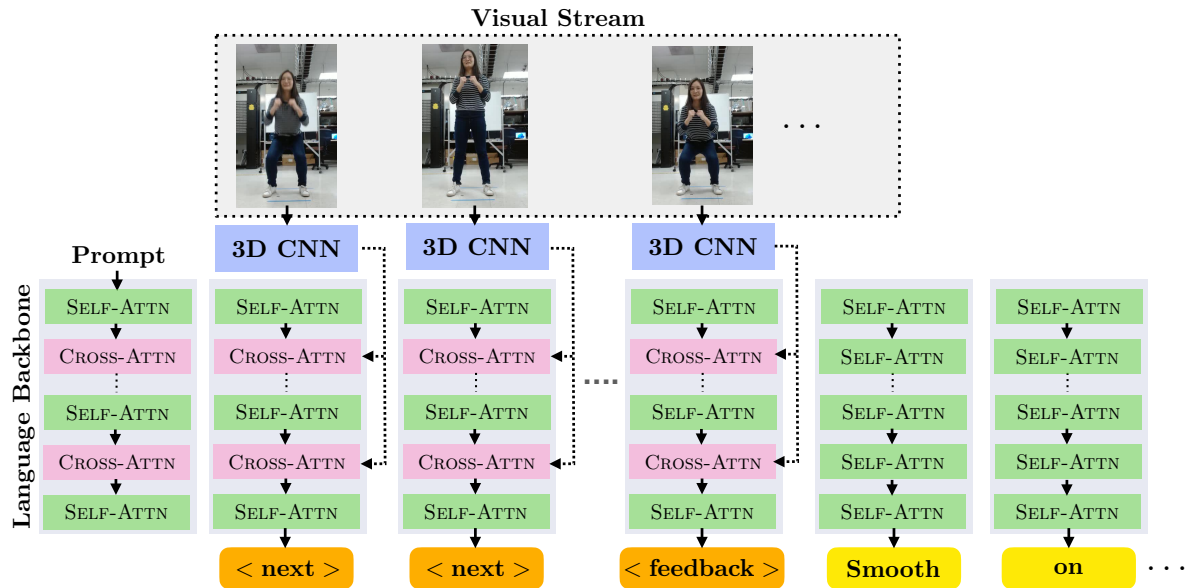


Figure 4. Architecture of our interactive Coach-LLaMA model. The visual stream is processed by a 3D CNN, the language backbone is a LLaMA-2-7B model and the special action tokens (<next> and <feedback>) are highlighted in orange.

	Short clips		Long-range	
	Train	Test	Train	Test <sup>†</sup>
Number of videos	290,775	16,429	153	69
Unique participants	1,800+	100	21	7
Average duration (s)	5.62	5.59	213.43	213.67
Exercises per video	1	1	5-6	5-6
Total number of exercises	148	148	23	23
Fitness Questions Dataset				
Total High-level Questions	1,193,056	78,390	-	-
Total Fine-grained Questions	404,082	80,694	-	-
Fitness Feedback Dataset				
Average feedback events per video	2.2	2.2	41.2	40.7
Average silence period (s) <sup>††</sup>	n/a	n/a	5.15	5.26
Average feedback length (words)			11	
Total feedbacks <sup>†††</sup>			8,000+	

Table 2. Dataset summary statistics. <sup>†</sup>The test split of the long-range videos forms our FIT-COACH benchmark. <sup>††</sup>Only a single feedback is provided at the *end* of the short clips. <sup>†††</sup>We provide a diverse set of alternative feedbacks that can be sampled for feedback events where applicable.

## B.2. Training Scheme

Our COACH-LLAMA model is trained end-to-end in the following three stages:

**Vision Backbone.** We first pre-train the vision backbone alone on ImageNet[8] followed by the short-clips video collection described in Appendix A, to help recognize fine-grained fitness activities including speed, range of motion, and common errors. In this pre-training stage, we train the model directly on the fine-grained labels available on our short-clips dataset. Further training details are provided in the supplemental.

**Short-clip videos.** Next, we train our full COACH-LLAMA model end-to-end on the fitness questions and fitness feedback annotations on the short-form clips from our FIT-COACH dataset. The purpose of this stage is to align the adapter and LM with the pre-trained action recognition capability of the vision backbone, and to link visual concepts to conversational-style language. We use LoRA [10] with  $r = 128$  for all key, value, and query matrices as well as the final LM decoding head. We train the model for 1 epoch with a learning rate of  $5 \times 10^{-6}$ .

**Long-range videos.** Finally, we train our model to generate timely feedback using the long-range videos from the FIT-COACH fitness feedbacks dataset. LoRA is also used in this stage with a smaller learning rate of  $1 \times 10^{-6}$ . Additional training details are provided in the supplemental.

**Dealing with uncertainty.** Our FIT-COACH benchmark includes situations where multiple relevant feedbacks are equally valid as described in Appendix A. Unlike turn-based models, our interactive COACH-LLAMA model can handle this uncertainty in a principled manner due to its end-to-end training and the use of action tokens. This allows us to use sampling techniques, *e.g.*, beam search, to

generate multiple random samples, reflecting the range of likely feedbacks. In Tab. 3, we demonstrate this ability using  $N = 3$  samples and a Best-of-N evaluation metric, where the best sample is selected based on the Temporal-F-score. This approach significantly boosts the performance of our COACH-LLAMA, proving its ability to effectively capture the distribution of possible feedbacks.

## B.3. Detailed Evaluation of Fitness Domain Knowledge

Next, we evaluate the state-of-the-art (turn-based) Video-LLaMA and Video-ChatGPT models on questions from our FIT-COACH dataset to test their fitness domain knowledge. We test the models on a random subset of 2000 questions from the fitness questions and on feedback generation on the short-clip videos in the fitness feedbacks dataset. For evaluation we use Mixtral-Instruct-0.1, which is prompted to score the answers from the fitness questions dataset on the basis of factual accuracy and on the relevance of predicted feedbacks in case of the feedbacks dataset. The scores are in the range of 1-5. We see in Tab. 4 that the off-the-shelf Video-LLaMA and Video-ChatGPT models perform poorly due to the lack of fitness domain knowledge. This is further illustrated in the qualitative examples in Fig. 6. The Video-LLaMA and Video-ChatGPT model is unable to discern fine-grained motion, *e.g.*, the motion of the legs of the participant in case of high knees (Fig. 6 top). The Video-LLaMA and Video-ChatGPT models are also unable to provide relevant feedbacks. The feedbacks tend to be descriptive instead of being actionable (Fig. 6 bottom).

## B.4. Effect of Weighting Action Tokens

During fine-tuning on the FIT-COACH long-range videos, the `<next>` special action tokens constitute a large propor-

Method	T-F-Score $\uparrow$	T-BERT $\uparrow$	T-Rouge-L $\uparrow$	Mixtral-Score $\uparrow$
COACH-LLAMA (w/o 3D CNN)	0.62	0.488	0.070	3.01
COACH-LLAMA (Ours)	<b>0.68</b>	<b>0.563</b>	<b>0.125</b>	<b>3.43</b>

Table 3. Evaluation on the FIT-COACH benchmark using Best-of-N samples (N=3). Samples are drawn using beam search.

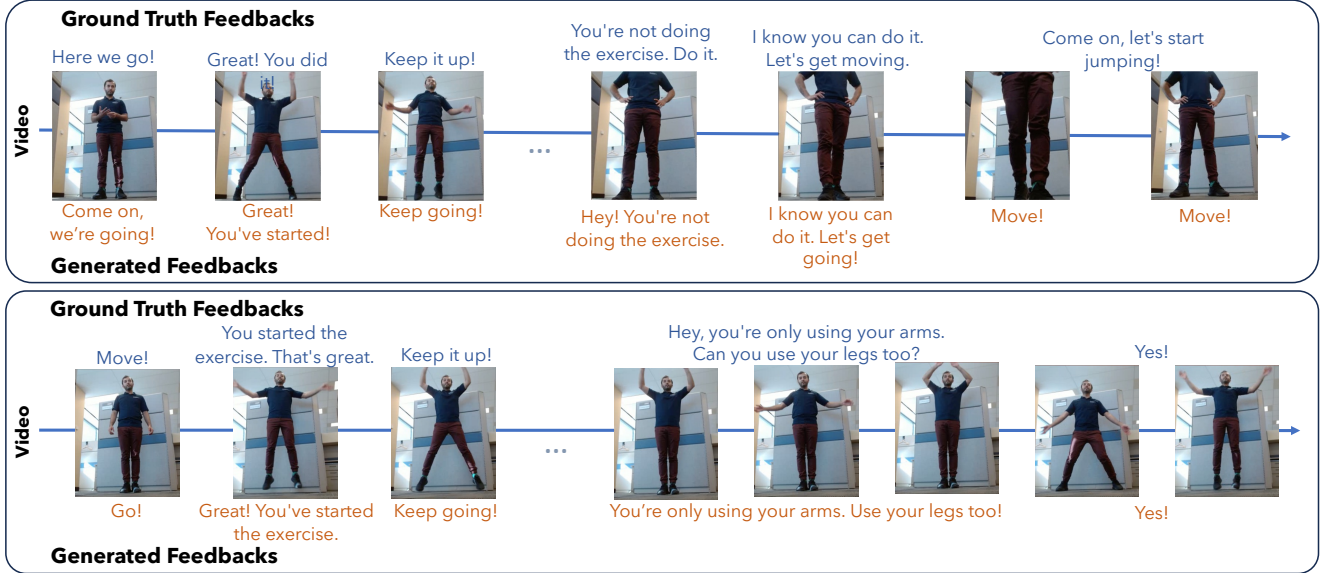


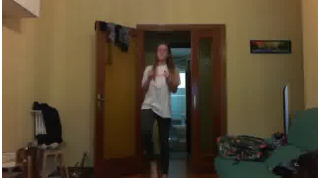
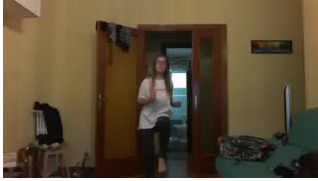
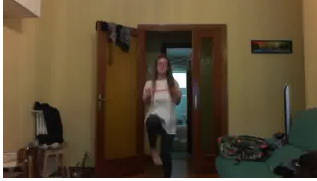

Figure 5. Feedbacks from our COACH-LLAMA model on the FIT-COACH benchmark.

tion of the tokens in training sequences as feedbacks are not present at every time-step. To address this, we weight the loss on the  $\langle \text{next} \rangle$  tokens during fine-tuning of our COACH-LLAMA model. We report effect of weighting the  $\langle \text{next} \rangle$  special action tokens during fine-tuning on the temporal-precision/recall/F-score in Fig. 7. We find that a weight of 0.1 works best. Higher weights leads to lower recall as the model prefers to observe the participant for longer, due to the large proportion of the  $\langle \text{next} \rangle$  token in training sequences.



Method	Factual-Accuracy $\uparrow$	Feedback-Relevance $\uparrow$
Video-LLaMA <sup>†</sup> [15]	1.71	1.34
Video-ChatGPT <sup>†</sup> [12]	1.80	1.42
COACH-LLAMA (Ours)	4.43	4.32

Table 4. Evaluation on fitness domain knowledge on our fitness questions and feedbacks dataset. (<sup>†</sup>indicates off-the-shelf models.)





**Question:** Did the user use only one leg?

**Video-LLaMA:** You are doing front lunges with the right leg, followed by a kick. ✗

**Video-ChatGPT:** Yes, the user is using only one leg to walk in the video. ✗

**COACH-LLAMA:** No, the user used both legs. ✓

---

**Question:** Provide an appropriate feedback for the user.

**Video-LLaMA:** We see a young man standing in a kitchen, wearing a red shirt and white shorts. ✗

**Video-ChatGPT:** The user has successfully demonstrated the ability to perform a balancing act on a pair of stools. ✗

**COACH-LLAMA:** This is awesome. Let's keep the intensity high! ✓

Figure 6. Evaluation on the fitness questions (top) and short-clip fitness feedbacks (bottom). State-of-the-art video-language models such as Video-LLaMA and Video-ChatGPT are unable to discern fine-grained motion or provide appropriate feedbacks.

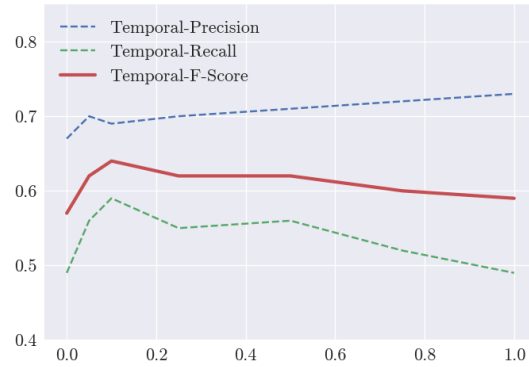


Figure 7. Effect of weighting the <next> action token on the Temporal-F-Score measured on the FIT-COACH benchmark.