# What to Say and When to Say it: Live Fitness Coaching as a Testbed for Situated Interaction

**Sunny Panchal**[1*] **Apratim Bhattacharyya**[1*] **Guillaume Berger**[1] **Antoine Mercier**[1]
**Cornelius Böhm**[2‡] **Florian Dietrichkeit**[‡] **Reza Pourreza**[1] **Xuanlin Li**[3§] **Pulkit Madan**[1]
**Mingu Lee**[1] **Mark Todorovich**[1] **Ingo Bax**[1] **Roland Memisevic**[1]
[1]Qualcomm AI Research[†] [2]Aignostics GmbH [3]UC San Diego

## Abstract

Vision-language models have shown impressive progress in recent years. However, existing models are largely limited to turn-based interactions, where each turn must be stepped (i.e., prompted) by the user. Open-ended, asynchronous interactions, where an AI model may proactively deliver timely responses or feedback based on the unfolding situation in real-time, are an open challenge. In this work, we present the QEVD benchmark and dataset, which explores human-AI interaction in the challenging, yet controlled, real-world domain of fitness coaching – a task which intrinsically requires monitoring live user activity and providing immediate feedback. The benchmark requires vision-language models to recognize complex human actions, identify possible mistakes, and provide appropriate feedback in real-time. Our experiments reveal the limitations of existing state-of-the-art vision-language models for such asynchronous situated interactions. Motivated by this, we propose a simple end-to-end streaming baseline that can respond asynchronously to human actions with appropriate feedback at the appropriate time.

## 1 Introduction

Datasets that combine visual information and language have greatly contributed to advancing the abilities of AI models over the past years, ranging from captioning [19], to visual questions answering [6], to visual dialogue [17], and beyond. Particularly impressive showcases of this progress are recent models such as GPT-4o [48] and Gemini [58], which can interact with users in real-time.

Despite the impressive recent progress, existing vision-language models still lag far behind human capabilities. While state-of-the-art models can be queried (e.g., through prompting) to comment on events shown in the camera stream, they lack the ability to interact asynchronously with the user as demanded by the situation, rather than only when prompted. Such interactive scenarios that are grounded in the spatial and temporal context of an unfolding situation are commonly referred to as "situated" [5, 12, 13]. Addressing such situated interactive scenarios will be a key to developing real-world assistive vision-language models.

A notable type of situated interaction is the instructional or coaching scenario, where an instructor guides a user through a complex activity, such as live fitness coaching. The real-world domain of

[*]Authors contributed equally
[†]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
[‡]Work performed at TwentyBN GmbH
[§]Work performed at Qualcomm AI Research

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-2415

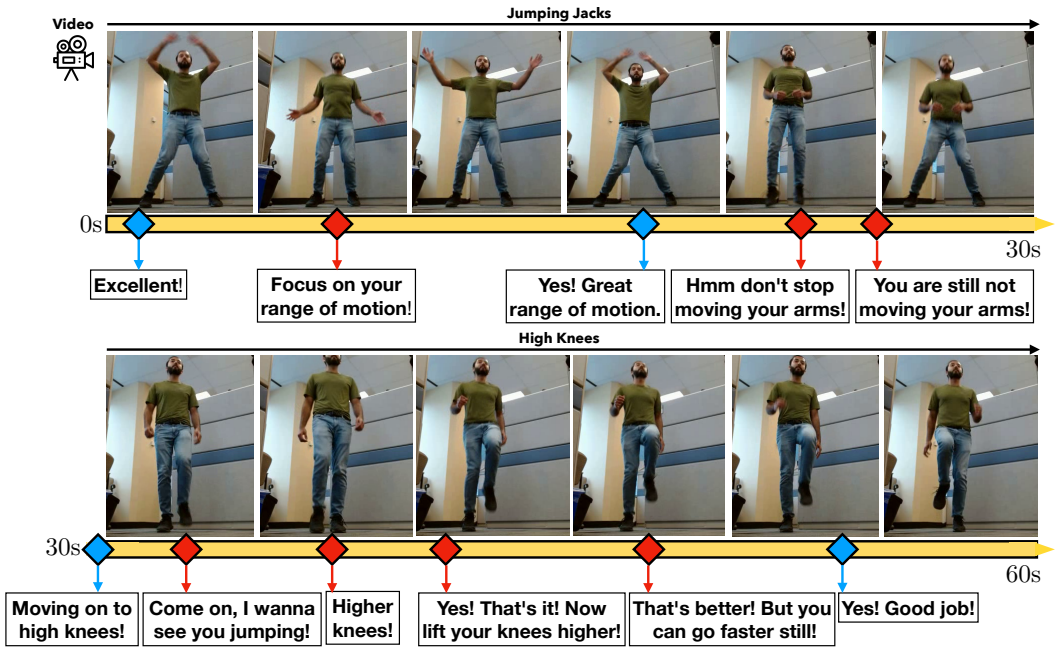



Figure 1: Long-range interactive videos from our QEVD-FIT-COACH benchmark. Live feedbacks provided to the participants are shown below each frame. Corrective feedbacks in red.

live fitness coaching has several benefits that make it an ideal test-bed for studying situated assistive vision-language models: Firstly, fitness routines have a controlled structure in that users are expected to, but may not, follow a prescribed series of actions. Secondly, despite the structured nature of fitness routines, coaching in this domain remains a complex and unsolved problem for current vision-language systems. The nuances of human motion pose a significant challenge for these systems to effectively understand the dynamic situation and respond interactively. Finally, live fitness coaching represents a rapidly growing real-world application. The increasing popularity of home workouts [2], highlights the practical need and potential impact of developing effective solutions in this area. A successful vision-language model for live fitness coaching could thus offer users tangible benefits.

Currently available large-scale video datasets [24, 27, 46, 55, 72] provide a rich set of annotations with expert demonstrations in domains such as cooking or house-hold activities. Expert demonstrations alone are not sufficient for real-world instructional scenarios, such as live fitness coaching, where the user may make mistakes. This has been addressed recently for some ego-centric instructional tasks by [10] and [62]. Successfully guiding a user through a fitness routine additionally necessitates the ability to understand fine-grained human actions, to provide appropriate instructions, and to provide corrective feedback—grounded in those fine-grained human actions—to correct any mistakes made by the user.

Overall, our main contributions are: (1) We propose the first large scale benchmark and dataset, Qualcomm Exercise Videos Dataset (QEVD), aimed at the development of video-language models for live coaching. QEVD* contains over 474+ hours of videos for fitness activity recognition and coaching. It includes short-clip videos (QEVD-FIT-300K) (∼5 seconds in length) annotated with 1M+ question-answer pairs, and long-range videos (>3 minutes in length) annotated with live feedback (QEVD-FIT-COACH, cf., Fig. 1); (2) We perform a comprehensive evaluation of state-of-the-art VLMs on the QEVD-FIT-COACH benchmark, revealing that the task is largely unsolved and offers significant room for improvement; (3) As a step towards closing the gap towards situated interaction, we propose a novel video-language model, STREAM-VLM, which, instead of being limited to turn-based interactions, can decide on-the-fly when and what to say to the user. The model is trained end-to-end to perform real-time visual interaction with a user based on live camera input.

*Data and code are available at `https://www.qualcomm.com/developer/software/qevd-dataset` and `https://github.com/Qualcomm-AI-research/FitCoach`

https://doi.org/10.52202/079017-2415

Table 1: Comparison of dataset statistics. *Domain*: target domain of the dataset; *Human Actions*: whether the dataset contains fine grained human actions; *Interactive*: whether the dataset captures interactions between two or more agents, e.g., a fitness coach and a participant; *Mistakes*: whether the dataset contains correct and incorrect actions towards a task; *Corrective Feedbacks*: whether the dataset contains corrective feedback provided in response to incorrect actions; *Domain Expertise*: whether the dataset contains the required domain expertise to provide fine-grained feedbacks for mistakes; *Length*: total length in hours.

| Dataset | Domain | Human Actions | Interactive | Mistakes | Corrective Feedbacks | Domain Expertise | Length |
|---|---|---|---|---|---|---|---|
| Action Recognition Datasets | | | | | | | |
| NTU RGB+D [51] | Fitness | ✓ | × | × | × | ✓ | – |
| FineGym [52] | Fitness | ✓ | × | × | × | ✓ | 708 |
| Procedural Activity Datasets | | | | | | | |
| YouCook2 [72] | Cooking | × | × | × | × | × | 176 |
| Epic-Kitchens [16] | Cooking | × | × | × | × | × | 100 |
| HowTo100M [46] | Daily-life | ✓ | × | × | × | × | 134k |
| Ego-4D [23] | Daily-life | × | × | × | × | × | 3670 |
| Ego-Exo4D [24] | Daily-life | × | × | ✓ | × | × | 1422 |
| Assembly-101 [50] | Toy assm. | × | × | ✓ | × | × | 513 |
| Interactive AI Assistant Datasets | | | | | | | |
| WTAG [10] | Cooking | × | ✓ | ✓ | ✓ | × | 10 |
| HoloAssist [62] | Obj. manip. | × | ✓ | ✓ | ✓ | × | 166 |
| QEVD (Ours) | Fitness | ✓ | ✓ | ✓ | ✓ | ✓ | 474 |

## 2 Related Work

**Datasets for Activity Recognition.** There is a large body of work on visual activity recognition. This includes NTU RGB+D 120 [51], FineGym [52], UCF101 [56], Kinetics [14], Moments in Time [47], ActivityNet [27], AVA-Kinetics [33], Charades [55], Something-Something [22], Something-Else [44], and others. Unlike our QEVD benchmark and dataset, these datasets do not contain detailed multi-modal annotations such as questions and feedbacks, or long videos with multiple actions or events. This restricts their utility in the development of interactive video-language models. Furthermore, while these datasets focus on a wide range of human activities, they contain only a few fitness activity related classes, if any (Tab. 1, col 2).

**Datasets for Procedural Activities.** The Epic-Kitchens [16] dataset provides ego-centric videos of non-scripted daily kitchen activities, with post-hoc recorded narrations. YouCook2 [72] provides cooking videos annotated with instructions on how to prepare specific meals, largely featuring expert chefs. HowTo100M [46] provides narrated instructional videos, featuring a variety of activities. These datasets are not interactive (Tab. 1, col 4) as they feature first-person instructions or narrations. Furthermore, they largely feature experts and do not include mistakes likely to be made by novices (Tab. 1, col 5). In contrast, QEVD is interactive, features participants with diverse skill levels, and thereby, includes mistakes likely to be made by novices.

Ego-Exo4D [24] includes a highly diverse set of activities, performed by participants with a variety of skill levels. Assembly-101 [50] features videos of people with diverse skill levels assembling and disassembling 101 "take-apart" toy vehicles. However, these datasets are not interactive and they do not include corrective feedbacks (Tab. 1, col 6). QEVD is interactive and includes corrective feedbacks from the perceptive of the fitness coach.

Similar to our work, WTAG [10] and HoloAssist [62] include corrective feedbacks and are focused on the development of interactive AI assistants. However, they focus on domains such as cooking or object manipulation. As they are recorded from an ego-centric perspective, they do not contain complex human actions (Tab. 1, col 3). Furthermore, while they include mistakes and associated corrective feedbacks, they do not include diverse examples of possible mistakes per target task (Tab. 1, col 7). QEVD is recorded from the perspective of a virtual fitness coach and includes fine-grained


https://doi.org/10.52202/079017-2415

Table 2: QEVD summary statistics. †The test split of the long-range videos forms our QEVD-FIT-COACH benchmark. †† Average is reported per exercise for the long-range videos. ††† Only a single feedback is provided at the *end* of the short clips.

| | QEVD-FIT-300K | | QEVD-FIT-COACH | |
|---|---|---|---|---|
| | **Train** | **Test** | **Train** | **Test**† |
| Number of Videos | 281,660 | 16,429 | 149 | 74 |
| Unique Participants | 1,800+ | 100 | 21 | 7 |
| Average Duration (s) | $5.6 \pm 1.1$ | $5.6 \pm 1.2$ | $213.4 \pm 3.1$ | $213.7 \pm 3.3$ |
| Exercises per Video | 1 | 1 | 5-6 | 5-6 |
| Total Number of Exercises | 148 | 148 | 23 | 23 |
| Total Classes | 1842 | 1558 | - | - |
| Fitness Questions | | | | |
| Total High-level Questions | 535,299 | 31,326 | - | - |
| Total Fine-grained Questions | 377,678 | 28,849 | - | - |
| Fitness Feedbacks | | | | |
| Total Feedbacks | 573,637 | 36,333 | 5,403 | 2,484 |
| Average Feedbacks per Video†† | $2.0 \pm 10.1$ | $2.1 \pm 10.2$ | $5.0 \pm 1.3$ | $5.0 \pm 1.2$ |
| Average Silence Period (s)††† | n/a | n/a | $5.2 \pm 1.4$ | $5.3 \pm 1.2$ |
| Average Feedback Length (words) | $8.9 \pm 5.1$ | $9.2 \pm 5.1$ | $6.3 \pm 3.8$ | $6.6 \pm 4.0$ |

human actions – diverse exercises and their variations including a wide diversity of mistakes per exercise.

**Datasets for VQA and Reasoning.** ActivityNet Captions [27], VATEX [61], TRECVID [7], HD-VILA [64], TGIF [35], WebVid [8], Charades [55], STAR [63] and AGQA [25] among others, focus largely on video captioning and question answering tasks. Finally, there exist a wide range of datasets on visual reasoning [6, 30, 49], including visual dialogue, for example [17], all of which, in contrast to our work are based on still images rather than videos. FIXMYPOSE [31] contains annotated instructions for pose correction but is limited to pairwise images and is activity-agnostic.

**Models for Situated Interactions.** There is also a growing body of work on enabling LMs to generally reason over visual input [4, 26, 42, 54, 57, 60, 65, 67, 70]. However, the existing models can answer only high-level questions about depicted scenes and objects. Models for exercise feedback are discussed in [20, 53] among others. However, such models are based on the recognition of whether or not an exercise was performed, or on counting repetitions, rather than providing interactive guidance and reasoning about the user's movements from a third-person viewpoint, which is the focus of this work.

## 3 Fitness Interactive Coaching Dataset and Benchmark

We now introduce QEVD, including QEVD-FIT-300K, and the QEVD-FIT-COACH dataset and benchmark, in detail. We begin with a detailed description of the QEVD-FIT-COACH benchmark, followed by additional details of the QEVD-FIT-300K and QEVD-FIT-COACH datasets.

### 3.1 QEVD-FIT-COACH Benchmark

The QEVD-FIT-COACH benchmark contains videos of participants performing a structured workout while receiving live feedback. These feedbacks may be corrective, affirmative, or informative, depending on user activity, to improve their form and pacing as they follow the workout using the temporal structure described below.

**Feedback Structure.** The feedbacks in the QEVD-FIT-COACH benchmark have the following structure: At the start of each exercise, acknowledging feedback is given once the user has started; otherwise, a reminder to do so is provided. A corrective feedback is provided as soon as a mistake is clearly visible. Similarly, when the user begins to correct their mistake, feedback is provided to acknowledge and guide the user to successfully correct the error. If the user is performing the exercise

correctly, feedback focuses on repetition counting. When repetition counts are not possible, such as with deltoid stretches, users receive positive, encouraging feedback, regularly with an average silence period of 5 seconds between successive feedback. Finally, at the end of each exercise, a feedback focused on the overall performance during that exercise is provided. This temporal structure ensures that feedbacks in the QEVD-FIT-COACH benchmark occur at predictable time-steps, aligned to visually salient moments. The annotated feedbacks of each video were verified by a second annotator.

An example from the QEVD-FIT-COACH benchmark, illustrating a trimmed workout session with two exercises—jumping jacks followed by high knees—is shown in Fig. 1. The segment begins with affirmative feedback to acknowledge the participant starting the exercise. Next, a series of feedbacks to correct user mistakes and affirm their compliance are provided. Initially, the participant exhibits a low range of motion and is asked to correct this. Once the range of motion improves, the participant receives encouraging feedback. However, they then stop moving their arms, incurring another corrective feedback. Note that this feedback considers their previous arm movements. The user then moves on to the high knees exercise after being requested to do so. Here, the user receives corrective feedback to raise their knees to the appropriate height and to improve their pace. The session ends with positive encouraging feedback acknowledging that the user has currently performed the exercise. These examples highlight the highly interactive coaching sessions in our QEVD-FIT-COACH benchmark and showcase the tight coupling between the participant actions and timely feedback.

**Statistics:** In total, the QEVD-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 arranged in 30 second segments. The overall list of 23 exercises is provided in the appendix. It includes a total of 7 unique participants with a cumulative recording length of $\sim$20 minutes to $\sim$1.5 hours.

### 3.2 Fitness Datasets for Training

Together, the QEVD-FIT-300K and QEVD-FIT-COACH datasets in QEVD are designed to instill domain understanding for fitness coaching and provide effective feedbacks during live coaching sessions. They consist of three annotation types (Fig. 2), which are described in detail below.

**Fine-grained Fitness Activity Labels.** This includes $460+$ hours of labeled (short) videos (QEVD-FIT-300K) crowd-sourced from over $1,900$ unique participants in the wild. They cover $148$ different exercises and their variations including: varied pacing, performing common mistakes, and modified form. Exercise variations were determined top-down through consultation with expert fitness instructors. Participants were then provided detailed instructions and an accompanying reference video to perform the exercises and their pre-determined variations. Approximately $10$ variations were collected per exercise. We show such variations for the push-ups exercise in Fig. 2 (top). Additionally, there are $49$ types of general activities such as "grabbing a towel" or "drinking from a bottle". Video lengths are in the 2 to 10 second range. There are approximately $3,500$ videos per exercise on average ($\sim$300k clips overall) and a total of $1,800+$ fine-grained classes capturing the exercise variations and general activities. Each video and corresponding labels were manually reviewed for correctness by at least one unique crowd-worker. These labels support training vision models for fine-grained understanding of human motion associated with fitness exercises.

**Fitness Questions.** In addition to the fine-grained labels, question-answer pairs querying video properties are provided for each fine-grained short video. This data can be used to provide further grounding of the LM's concepts in the observed visual input. The questions can be broadly divided into two types: high-level and fine-grained, as shown in Fig. 2. The high-level questions are directed at the overall exercise type and performance of the participant, e.g., in Fig. 2, 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., "Was the user's squat shallow and used only one arm?", or "Is the user going as fast as possible?". These questions are generated using the Mixtral-8B-Instruct LLM [29] and the question generation process follows the scheme detailed in [38]. Overall statistics for this subset, including a breakdown of the provided splits, can be found in Tab. 2.


https://doi.org/10.52202/079017-2415


Figure 2: Example annotations available on the short video clips from the QEVD-FIT-300K dataset. Annotations include question/answer pairs from our fitness questions dataset and feedback from a coaching perspective.

**Fitness Feedbacks.** We provide a set of annotations from a second-person perspective to support feedback in a live coaching session. It contains both fine-grained short videos mentioned above (QEVD-FIT-300K) and additional long-range videos (QEVD-FIT-COACH). For the fine-grained short videos, an average of 2 feedbacks per video are provided (Tab. 2). These feedbacks occur at the end of each video, as shown in Fig. 2 (bottom). In the shown example, the feedbacks focus on improving the form of the participant, specifically, by encouraging them to squat deeper and punch with both arms. We collected an additional $\sim$9 hours of fitness coaching sessions following the same methodology as the QEVD-FIT-COACH benchmark. These sessions contain an average of 5 feedbacks per exercise, totaling approximately 35 feedbacks per workout, including instructions (See Tab. 2).

## 4 Baseline STREAM-VLM

Current state-of-the-art vision-language models [15, 34, 36, 43, 66, 69] are largely *turn-based*—they take an image or video as input along with an instruction and produce a textual output. In contrast, our QEVD-FIT-COACH benchmark requires models to provide feedback proactively, i.e., without explicit prompting, based solely on the participants' actions in a *streamed* setting. To this end, we propose a baseline streaming video language model, STREAM-VLM, specialized to the fitness coaching domain. It consists of a 3D-CNN-based vision backbone [1, 45] for understanding fine-grained fitness actions and a LLaMA-2-based language backbone [59]. Special action tokens are

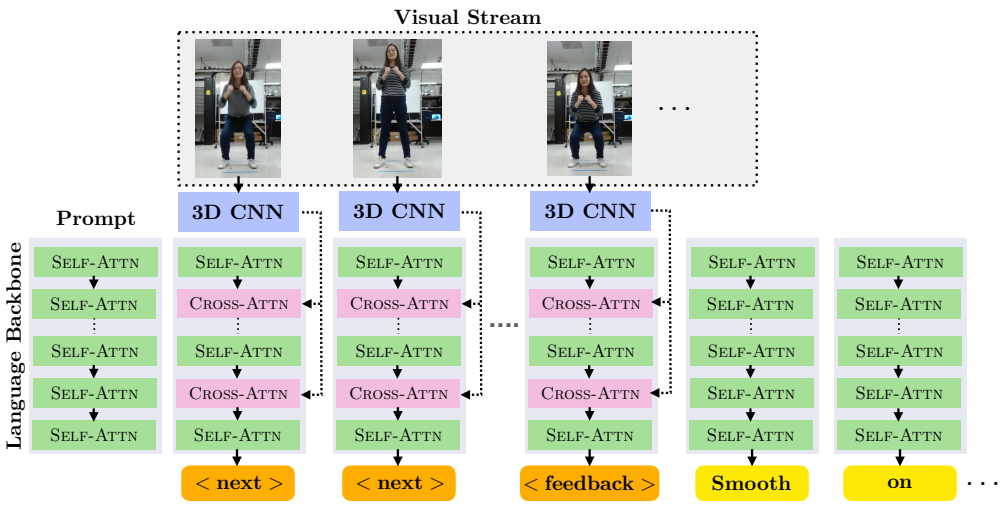



Figure 3: Architecture of the STREAM-VLM model. The visual stream is processed by a 3D CNN and the language backbone is a LLaMA-2-7B model; special action tokens (<*next*> and <*feedback*> ) are highlighted in orange.

introduced to enable feedback delivery without explicit prompting. We begin by describing the vision backbone in detail, followed by the special action tokens, and finally, the training scheme.

**Vision backbone.** Our vision backbone is designed to robustly recognize motion cues crucial for fitness coaching. This is in contrast to current state-of-the-art video-language models [43, 66], which typically use CLIP-/ViT-based vision encoders to capture scene content rather than motion. Specifically, our architecture is based on a publicly available 3D CNN [1, 45] capable of recognizing 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 both motion and content information of individual frames to make predictions—both of which are relevant to provide appropriate feedbacks. Additionally, the convolutional layers are causal, making the model well-suited for the streaming setting of our QEVD-FIT-COACH benchmark. Features from the vision backbone are fused into the LM backbone through cross attention at several layers following the methodology of [4, 11].

**Action tokens.** The STREAM-VLM baseline uses two special action tokens <*next*> and <*feedback*> to enable proactive feedbacks. The <*next*> token allows the model to opt not to say anything and request the next video frame as input from the visual stream. Conversely, when the model does decide to say something, it generates a <*feedback*> token. Through the introduction of these tokens, the model can be trained end-to-end to switch between stream observation and response generation without the need for external prompting heuristics. Specifically, in the coaching setting this allows the model to observe user activity and learn *when* to provide feedback based on *what* the user was observed doing.

In Fig. 3 the STREAM-VLM 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.

**Training scheme.** Our STREAM-VLM streaming baseline is trained end-to-end in three stages: (1) The vision backbone (3D CNN) is pre-trained on ImageNet[18], followed by the QEVD-FIT-300K short-clips video collection described in Sec. 3.2; (2) Next, the model is trained end-to-end on the fitness questions and fitness feedbacks annotations (excluding the long-range videos) from the QEVD-FIT-COACH dataset. The purpose of this stage is to align the vision backbone (3D CNN) and LM with the pre-trained action recognition capability of the vision backbone. Hence, only the adapter (cross-attention layer) weights are updated; (3) Finally, the model is fine-tuned on long-range videos from the QEVD-FIT-COACH fitness feedbacks subset with feedback annotations interleaved


https://doi.org/10.52202/079017-2415

Table 3: Zero-shot evaluation on the QEVD-FIT-COACH benchmark.

| Method | METEOR↑ | ROUGE-L↑ | BERT↑ | LLM-Acc.↑ |
|---|---|---|---|---|
| InstructBLIP [15] | 0.047 | 0.040 | 0.839 | 1.56 |
| Video-LLaVA [36] | 0.057 | 0.025 | 0.847 | 2.16 |
| Video-ChatGPT [43] | 0.098 | 0.078 | 0.850 | 1.91 |
| Video-LLaMA [66] | 0.101 | 0.077 | 0.859 | 1.29 |
| LLaMA-VID [34] | 0.100 | **0.079** | **0.859** | 2.20 |
| LLaVA-NeXT [69] | **0.104** | 0.078 | 0.858 | **2.27** |

Table 4: Evaluation of models fine-tuned on QEVD on the QEVD-FIT-COACH benchmark. ($^{\dagger}$indicates results of non-interactive models evaluated at regular intervals; $^{*}$indicates models fine-tuned by ourselves.)

| Method | METEOR↑ | ROUGE-L↑ | BERT↑ | LLM-Acc.↑ | T-F-Score↑ |
|---|---|---|---|---|---|
| Socratic-LLaMA-2-7B | 0.094 | 0.071 | 0.860 | 2.17 | 0.50$^{\dagger}$ |
| Video-ChatGPT [43]$^{*}$ | 0.108 | 0.093 | **0.863** | 2.33 | 0.50$^{\dagger}$ |
| LLaMA-VID [34]$^{*}$ | 0.106 | 0.090 | 0.860 | 2.30 | 0.50$^{\dagger}$ |
| STREAM-VLM | **0.127** | **0.112** | **0.863** | **2.45** | **0.56** |
| STREAM-VLM (w/o 3D CNN) | 0.090 | 0.083 | 0.857 | 2.11 | 0.51 |
| STREAM-VLM (w/o Pre-training) | 0.095 | 0.087 | 0.858 | 2.08 | 0.52 |
| STREAM-VLM (w/o Action-Tokens) | 0.125 | 0.110 | 0.861 | 2.41 | 0.50$^{\dagger}$ |

to reflect an interactive streaming setting. We limit the model training to 30-second individual exercise segments and leave it to future work to train on workouts spanning multiple exercises. The LLaMA-2 language backbone is fine-tuned using LoRA (dim = 32) [28]. The 3D CNN and adapter (cross-attention layer) weights are kept frozen. Additional details are provided in the appendix.

## 5 Experiments

In this section we evaluate current state-of-the-art (open source) video-language models and explore their limitations in the interactive streaming setting of our QEVD-FIT-COACH benchmark. We also evaluate our STREAM-VLM model and highlight potential avenues to address the key challenges associated with the QEVD-FIT-COACH benchmark.

### 5.1 Evaluation Metrics.

The following metrics are used to capture both the fluency ("what to say") and temporal accuracy ("when to say it") of generated feedback.

**Fluency.** We use the METEOR [9], ROUGE-L [37] and BERT [68] metrics to evaluate fluency. The METEOR, and ROUGE-L metrics assess lexical similarity between the ground truth and predicted feedbacks: e.g., in the case of a corrective feedback where the person is not moving their arms, these metrics would prefer predicted feedbacks referring to the "arm" and "not moving". The BERT score on the other hand matches feedbacks at a semantic level.

To compute these metrics, we first temporally match predicted and ground truth feedbacks. Each ground truth response is matched to the closest predicted response within a 3 second window, maintaining their temporal order. The respective METEOR, ROUGE-L and BERT scores are then computed on only the matched feedbacks.

**Automated evaluation (LLM-Accuracy).** In addition to the metrics above, we employ an LLM for holistic feedback evaluation [40, 43]. In contrast to the metrics above, LLMs offer the advantage that they better correlate with human preferences [71]. We provide the LM with ground truth and predicted feedbacks. The LM then scores the predicted feedbacks holistically for accuracy. We use the state-of-the-art open-source LLaMA-3-70B-Instruct [3] LLM, with scores in the range 1 to 5. This LLM-Accuracy metric along with METEOR, ROUGE-L and BERT metrics ensures that the

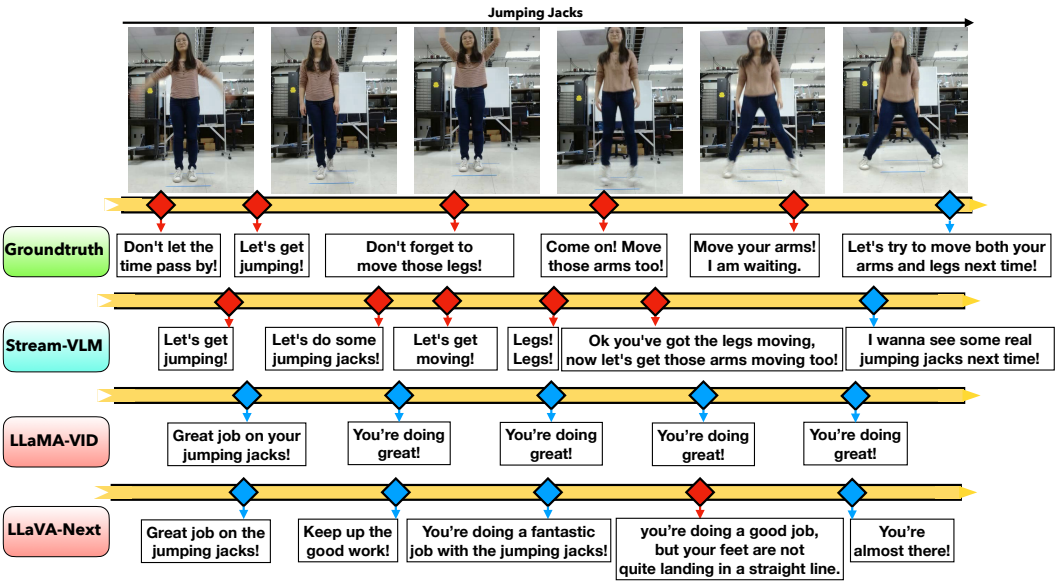



Figure 4: Predicted feedbacks on the FIT-COACH benchmark. The "turn-based" LLaMA-VID and LLaVA-NeXT models are unable to provide corrective feedback and instead generate overly generic and repetitive feedback. The STREAM-VLM model has learned to provide relevant feedback at the appropriate time.

predicted feedbacks match the ground-truth both at the semantic level while containing references to specific important terms.

**Temporal F-Score.** To measure temporal accuracy we assess whether predicted responses occur at the correct time-step and compute a temporal F-score. Predicted responses are classified as true or false positives based on whether they temporally match ground truth responses as described for the fluency metrics. Predicted responses without a ground truth match are false positives and ground truth responses without a matching predicted response are false negatives. This allows us to calculate temporal precision, recall, and hence, the temporal F-Score.

### 5.2 Evaluation on the QEVD-FIT-COACH Benchmark

We begin by evaluating state-of-the-art (open source) video-language models, including InstructBLIP [15], Video-LLaVA [36], Video-ChatGPT [43], Video-LLaMA [66], LLaMA-VID [34], and LLaVA-NeXT[39], on the QEVD-FIT-COACH benchmark (see Tab. 3). Since these models are "turn-based"—they cannot respond interactively to an input video—we prompt them to provide feedbacks at regular intervals. Specifically, to remain faithful to the streaming setting in our QEVD-FIT-COACH benchmark, we always prompt these models (zero-shot) with the entire video, including a history of generated feedbacks, up to the latest time-step. We use an interval of 5 seconds, equivalent to the average silence period within an exercise segment in the QEVD-FIT-COACH dataset. While LLaMA-VID [34] and LLaVA-NeXT [39] perform best among zero-shot baselines, overall performance across all zero-shot models is weak as shown in Tab. 3. We present qualitative examples in Fig. 4, highlighting the repetitive and uninformative nature of the feedback provided by LLaMA-VID and LLaVA-NeXT. They are unable to provide corrective feedback at the right time largely due to their lack of fitness domain knowledge and their "turn-based" nature.

Next, we address the lack of fitness domain knowledge by fine-tuning Video-ChatGPT and LLaMA-VID on QEVD following the process discussed in Sec. 4. As shown in Tab. 4, fine-tuning significantly improves performance as expected. However, the performance gain is still limited by the CLIP-/ViT-based visual encoders, which are not well-suited for representing fine-grained human motion, not to mention the limitations incurred from their turn-based nature.

To deal with these issues, our STREAM-VLM baseline uses a 3D CNN trained to recognize fine-grained fitness activities and special action tokens to enable interactive feedback. We also consider the following ablations of the model: (1) instead of the 3D CNN, we use a CLIP-based encoder sim-


https://doi.org/10.52202/079017-2415

ilar to Video-ChatGPT [43] ("w/o 3D CNN"); (2) we skip pre-training the STREAM-VLM with the QEVD-FIT-300K short-clips fitness questions and feedbacks dataset ("w/o Pre-training"); (3) we use a non-interactive turn-based version without the <*next*> and <*feedback*> action tokens ("w/o Action Tokens"). We also consider a text-only Socratic model, Socratic-LLaMA-2-7B [65]. In this model, we prompt the language-only LLaMA-2-7B LLM to generate feedback for the previous 5 seconds of user activity, provided as a list of activity descriptions in 1-second intervals. Similar to the zero-shot video-language model evaluations, the full history of described activity and generated feedbacks is included in the prompt. The textual description of user activity is based on the activations of the aforementioned fine-tuned 3D CNN. An example prompt is provided in the appendix.

The results in Tab. 4 demonstrate that the STREAM-VLM model surpasses the performance of the other models. Crucially, we see a significant improvement in the temporal F-score (0.59 vs 0.50) in comparison to the turn-based models. This is also illustrated in Fig. 4, where the STREAM-VLM model is shown to provide relevant corrective feedback at the appropriate time as opposed to the turn-based baselines. The improved quality of the feedbacks is also reflected in the METEOR, ROUGE-L and LLM-Accuracy metrics. The drop in both fluency and temporal accuracy resulting from the pre-training ablation (w/o Pre-training) supports the quality and utility of the fitness feedbacks and questions within our QEVD-FIT-300K dataset. Furthermore, the advantage of the 3D CNN is demonstrated by two observations: the weak performance of STREAM-VLM when the 3D CNN is ablated (w/o 3D CNN), and the strong performance of the Socratic-LLaMA-2-7B baseline. In the latter, an off-the-shelf LLaMA-2-7B model outperforms state-of-the-art vision-language models using only the activations of the 3D CNN as a prompt.

Overall, while there is significant room for improvement, these results suggest that end-to-end training is a viable path towards good performance on our QEVD-FIT-COACH benchmark and more broadly, on the task of responding interactively to events within a visual stream.

## 6 Conclusion

We propose QEVD, a novel interactive visual coaching benchmark and dataset, as a test-bed for real-time, real-world situated interaction, and demonstrate that this task is challenging for existing LLM-based architectures. As a first step towards closing the gap to situated interaction, we present STREAM-VLM, a streaming vision-language model baseline that learns not only what to say, but also when to say it, based on user activity in the incoming video stream. Overall, we consider our work a starting point for research into end-to-end training of domain-specific interactive vision models, and hope that our data and baselines will encourage further work in this area.

**Privacy and ethics.** The data was collected under a direct agreement with the crowd workers, permitting research and commercial use. Furthermore, a detector was used on all videos to detect any issues, such as individuals in the background, followed by manual inspection of the videos that scored above a threshold, to remove such videos from the collection. Personal identifiable information from the videos was removed to the extent possible, e.g., audio and meta-data. Participants received appropriate and fair compensation for the regions where they were located.

**Limitations.** Our work shows that contextual, situated interactions are possible to a degree for an AI model, when a significant amount of aligned training data is made available, and the interaction is confined to a highly restricted (albeit real-world) task domain. The ability to interact in broader domains, with less domain-specific training data, and with higher accuracy are open research problems. A related open problem is supplementing visual real-time input with speech input. A further limitation is that the predictions of models trained on the data cannot be guaranteed to be free from any bias with respect to, for example, a subject's age or gender.

**Broader Impact.** In addition to potential bias mentioned in the previous paragraph, language models can produce harmful and biased content, make incorrect claims and produce wrongful advice. This needs to be taken into account when interacting with, deploying or building on these models, particularly in sensitive domains like fitness coaching where incorrect advice may lead to physical harm. Although the grounding in visual input supports the generation of language that is contextual, it is not a remedy against these deficiencies of language models. It is also important to consider that any computer vision model processing visual information about human subjects could, in principle, extract information beyond what is required for the use-case, such as biometric information.

## References


[1] Sense Core. https://github.com/quic/sense. [Online; accessed Oct-2024].

[2] Statista home fitness. https://www.statista.com/topics/12564/home-fitness/#topicOverview. [Online; accessed Oct-2024].

[3] AI@Meta. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md. [Online; accessed Oct-2024].

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022.

[5] Prithviraj Ammanabrolu, Renee Jia, and Mark O Riedl. Situated dialogue learning through procedural environment generation. In *ACL*, 2022.

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015.

[7] George Awad, Asad A. Butt, Keith Curtis, Yooyoung Lee, Jonathan G. Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, João Magalhães, David Semedo, and Saverio G. Blasi. TRECVID 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search. In *TREC Video Retrieval Evaluation*, 2018.

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021.

[9] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005.

[10] Yuwei Bao, Keunwoo Peter Yu, Yichi Zhang, Shane Storks, Itamar Bar-Yossef, Alexander De La Iglesia, Megan Su, Xiao-Lin Zheng, and Joyce Chai. Can foundation models watch, talk and guide you step by step to make a cake? In *EMNLP Findings*, 2023.

[11] Apratim Bhattacharyya, Sunny Panchal, Mingu Lee, Reza Pourreza, Pulkit Madan, and Roland Memisevic. Look, remember and reason: Visual reasoning with grounded rationales. In *ICLR*, 2024.

[12] 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.

[13] Rodney A. Brooks. Intelligence without reason. In *IJCAI*, 1991.

[14] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[15] 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.

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. In *IJCV*, 2022.

[17] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *CVPR*, 2017.


 https://doi.org/10.52202/079017-2415

[18] 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.

[19] 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.

[20] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *CVPR*, 2021.

[21] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. In *Communications of the ACM*, 2021.

[22] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017.

[23] 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, et al. Ego4d: Around the world in 3, 000 hours of egocentric video. In *CVPR*, 2022.

[24] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. *CoRR*, abs/2311.18259, 2023.

[25] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA: A benchmark for compositional spatio-temporal reasoning. In *CVPR*, 2021.

[26] Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. *CoRR*, abs/2211.11559, 2022.

[27] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[28] 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.

[29] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts. *CoRR*, abs/2401.04088, 2024.

[30] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

[31] Hyounghun Kim, Abhay Zala, Graham Burri, and Mohit Bansal. FIXMYPOSE: pose correctional captioning and retrieval. In *AAAI*, 2021.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *CoRR*, abs/2005.00214, 2020.

[34] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *CoRR*, abs/2311.17043, 2023.

[35] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, 2016.

[36] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *CoRR*, abs/2311.10122, 2023.

[37] Chin-Yew Lin and Franz Josef Och. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, 2004.

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[39] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

[40] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *EMNLP*, 2023.

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[42] Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *CoRR*, abs/2304.09842, 2023.

[43] 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.

[44] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*, 2020.

[45] Antoine Mercier, Guillaume Berger, Sunny Panchal, Florian Dietrichkeit, Cornelius Böhm, Ingo Bax, and Roland Memisevic. Is end-to-end learning enough for fitness activity recognition? *CoRR*, abs/2305.08191, 2023.

[46] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[47] Mathew Monfort, Carl Vondrick, Aude Oliva, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa M. Brown, Quanfu Fan, and Dan Gutfreund. Moments in time dataset: One million videos for event understanding. In *IEEE PAMI*, 2020.

[48] OpenAI. "Hello gpt-4o.". `https://openai.com/index/hello-gpt-4oâĄă`. [Online; accessed Oct-2024].

[49] Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301, 2022.

[50] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, 2022.

[51] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

 https://doi.org/10.52202/079017-2415

[52] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020.

[53] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *CVPR*, 2020.

[54] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving AI tasks with chatgpt and its friends in huggingface. *CoRR*, abs/2303.17580, 2023.

[55] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016.

[56] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012.

[57] Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. *CoRR*, abs/2303.08128, 2023.

[58] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023.

[59] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023.

[60] Junke Wang, Dongdong Chen, Chong Luo, Xiyang Dai, Lu Yuan, Zuxuan Wu, and Yu-Gang Jiang. Chatvideo: A tracklet-centric multimodal and versatile video understanding system. *CoRR*, abs/2304.14407, 2023.

[61] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, 2019.

[62] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, and Marc Pollefeys. Holoassist: an egocentric human interaction dataset for interactive AI assistants in the real world. In *ICCV*, 2023.

[63] Bo Wu, Shoubin Yu, Zhenfang Chen, Josh Tenenbaum, and Chuang Gan. STAR: A benchmark for situated reasoning in real-world videos. In *NeurIPS*, 2021.

[64] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, 2022.

[65] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek 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.

[66] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *EMNLP - System Demonstrations*, 2023.

[67] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *CoRR*, abs/2303.16199, 2023.

[68] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020.

[69] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.

[70] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *CoRR*, abs/2302.00923, 2023.

[71] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

[72] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018.

 https://doi.org/10.52202/079017-2415

# Paper Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] , the code for the baseline models are from publicly available sources and appropriate instructions are provided.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Existing models are evaluated using the publicly available checkpoints; for the Stream VLM we report mean scores and variance is small (as can be seen from the multiple different results across the different ablations)
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [Yes]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? All participants received appropriate and fair compensation for the regions where the participants were located.

# Appendix

## A Overview

Here we provide: (1) Additional qualitative examples from QEVD-FIT-300K and QEVD-FIT–COACH; (2) Additional data collection and annotation details, including instructions provided to the crowd-workers; (3) Additional training details of the STREAM-VLM model; (4) Details of the prompts used for the zero-shot baselines in Tab. 3, including the prompt used to compute the LLM-Accuracy described in Sec. 5.1.

## B Additional Qualitative Examples from QEVD-FIT-300K

In Fig. 5, we provide additional annotation examples from the fitness questions and fitness feedbacks on short-clip videos. As shown, high-level questions focus on overall aspects of the video, e.g., "What exercise is the user doing? Describe how they are doing it.". On the other hand, fine-grained questions focus on details such as the user's rate of punch (Fig. 5 top) or whether the user's back is rounded (Fig. 5 bottom). The feedbacks on these short-clips provide positive reinforcement, e.g., "This is looking great!" if the user is doing the exercise correctly (Fig. 5 top) and also provide useful advice, e.g., toe alignment while performing squats (Fig. 5 bottom).

## C Additional Data Collection Details

Additional details pertaining to the various subsets of QEVD are provided in this section.

### C.1 QEVD-FIT-300K

Here we explore the labels associated with the short video clips in the QEVD-FIT-300K dataset. These labels are crucial for generating fitness questions and providing feedback for short-clip videos.

**Video collection.** The short-clip exercise videos in the QEVD-FIT-300K dataset were collected through a simple web interface as shown in Fig. 8. The crowdworkers were shown a series of demonstration videos for the target exercise and were asked to replicate them. The exercises and their fine-grained variants were pre-determined in consultation with fitness coaches.

**Coarse and fine labels.** As described in Sec. 3.2 in the main paper, the short video clips are annotated with coarse labels and fine-grained attributes in addition to the questions and feedbacks. Labels were verified for correctness by at least one additional human annotator. The full list of coarse labels and fine-grained attributes can be found at `https://developer.qualcomm.com/software/ai-datasets/qevd`.

**Quantitative labels.** For a subset of exercises, additional quantitative properties like speed and range of motion can be derived through simple pairwise comparative annotation campaigns. We asked crowd-workers to compare and rank randomly selected pairs of clips from 33 exercises. The ranking allowed us to determine an ordering within an exercise. Finally, by quantizing the targets into a set of distinct groups we derived additional fine-grained labels. Such fine-grained labels would allow downstream applications to provide more precise feedback.

**Feedback Labels.** Additionally, we collect feedbacks for the short videos from the perspective of a fitness coach. These feedbacks correspond to the fine-grained events for 136 exercises, totaling over 8500 unique feedbacks.

**High-level Fitness Questions.** As described in the main paper, the fitness questions dataset consists of multiple question-answer pairs per short-clip video (Figs. 2 and 5) which are generated semi-automatically using the coarse-grained labels, fine-grained attributes and quantitative labels. We convert these raw labels and attributes to conversational-style questions answer pairs using Mixtral [29]. An example prompt for the high-level question "What exercise is the user doing?":


https://doi.org/10.52202/079017-2415

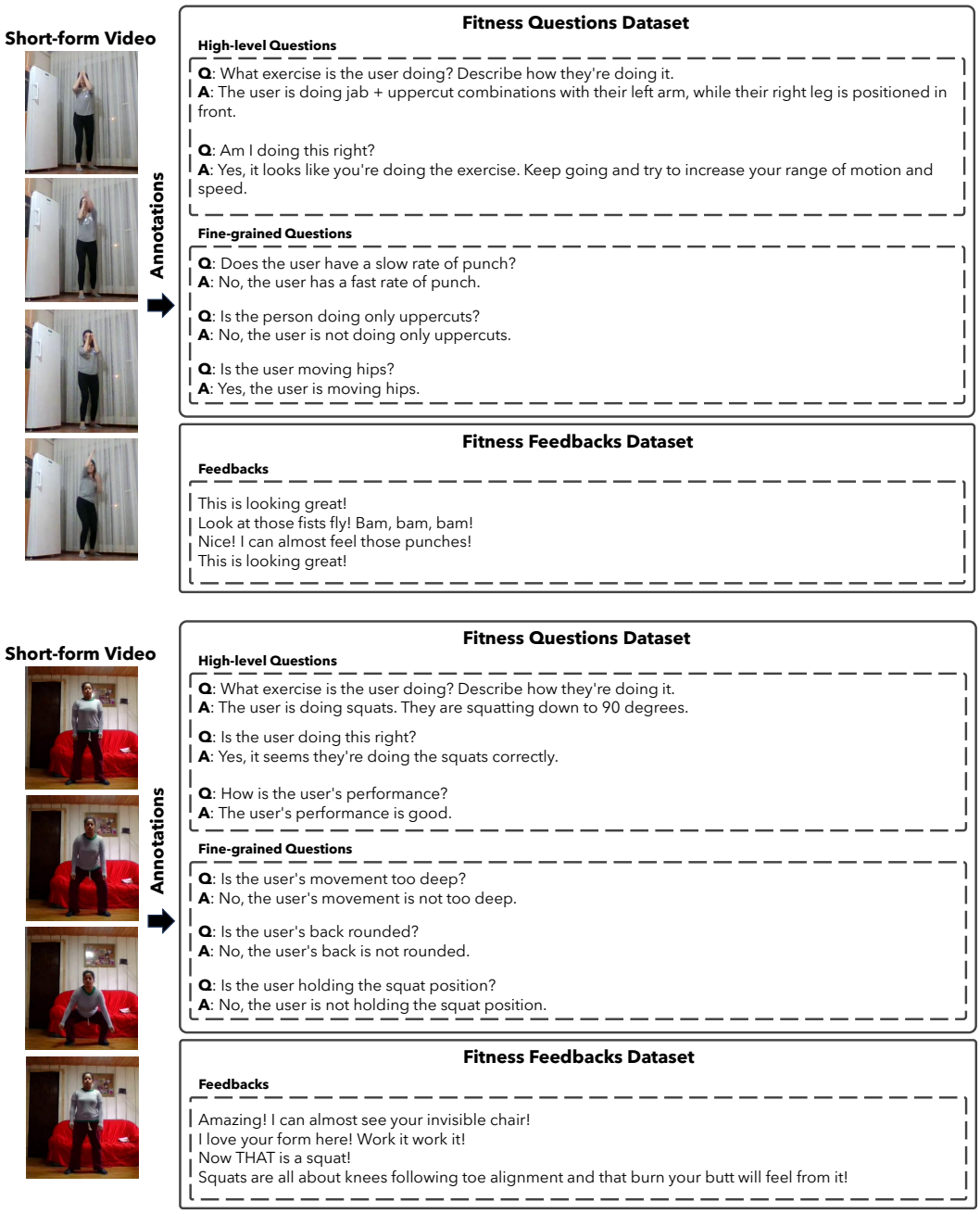


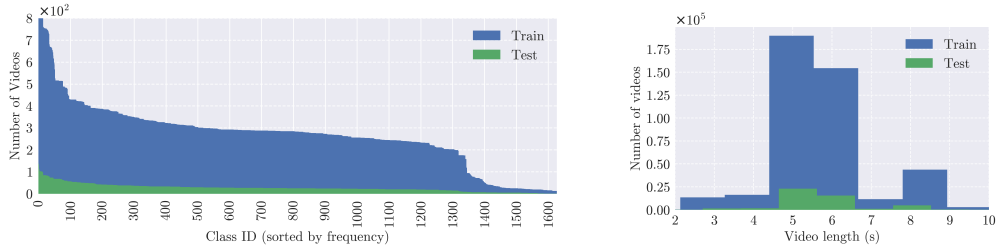

Figure 5: Additional example annotations available on the short video clips from the QEVD-FIT-300K dataset (see also Figs. 1 and 2 in the main paper).

Identify the exercise being performed from the provided templated exercise names and provide the answer in response to the question "What exercise is the user doing?".

The provided names will be in one of these formats (description after '–' not part of the template):
1) <exercise_name> (<variant>) – where variant may describe the side of the body being used or describing the position of the body.
2) <exercise_name>

For example:
Templated name: spider man pushup
Descriptive response: The user is doing spider man pushups.
Templated name: lunges (left leg out in front)
Descriptive response: The user is doing lunges with their left leg in front.

https://doi.org/10.52202/079017-2415




(a) Distribution of fine-grained class labels. (b) Distribution of clip lengths.

Figure 6: Statistics of the short-clips in the QEVD-FIT-300K dataset.

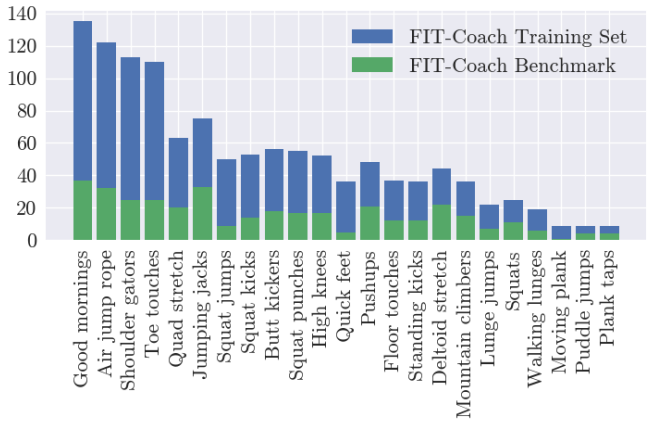



Figure 7: Exercise distribution in (long-form) videos from the QEVD-FIT-COACH benchmark and dataset splits.

Give a descriptive response for the following:
Templated name: <templated name>
Don't provide an explanation.

where, <templated name> is filled in with each query label.

**Fine-grained Fitness Questions.** The fine-grained questions per short-clip video (Figs. 2 and 5) are generated using the annotated fine-grained attributes, quantitative labels and feedbacks. We convert these raw labels and attributes to conversational-style questions answer pairs using Mixtral [29].

### C.2 QEVD-FIT-COACH

The set of exercises included in the QEVD-FIT-COACH benchmark and dataset (long-range videos) is shown in Tab. 5. The 23 unique exercises are highly diverse, of varying difficulty levels and require a wide variety of motion types by the participants. Note that this set of exercises is a subset of all exercises included in the QEVD-FIT-300K dataset. We show the distribution of these exercises across splits in Fig. 7.

**Video Collection.** Long-range workout videos were collected by letting users perform workout sessions consisting of three 1-minute sections, where each section consisted of two exercises. Candidate exercises per section are shown in Tab. 5. Users were instructed to perform specific exercises for the indicated duration while facing the camera to imitate a virtual fitness coaching setup. They were also instructed to perform common mistakes and their corrections (similar to the short-clips). Temporally aligned textual coaching instructions were subsequently cleaned-up and verified using a simple web interface. All feedbacks were reviewed by at least one additional annotator.


https://doi.org/10.52202/079017-2415

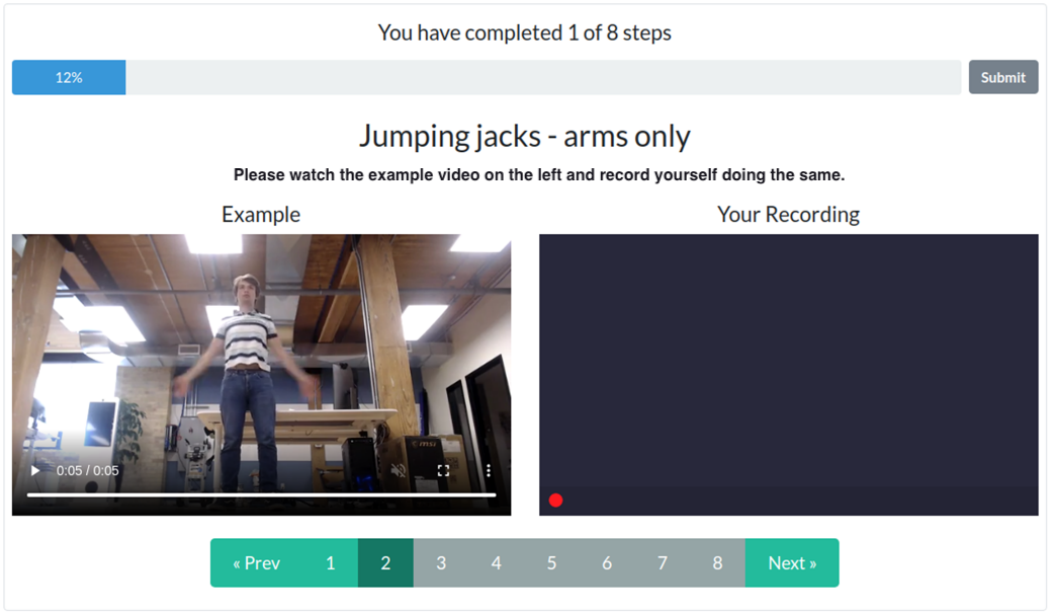



Figure 8: Simple QEVD-FIT-300K data collection web interface.

## D Training Details of the STREAM-VLM Model

Here, we provide additional details of training the STREAM-VLM model described in Sec. 4 in the main paper.

### D.1 Vision Backbone

After pre-training the 3D CNN model on ImageNet [18], we fine-tune the 3D CNN model on the fine-grained labels available on the short-clips dataset (full list in Appendix C.1). The 3D CNN model was trained for 500 epochs with a batch size of 32, a learning rate of $1 \times 10^{-4}$, an Adam optimizer [32, 41], and a frame-wise cross entropy loss objective.

**Dataset pre-processing:** The 3D CNN vision backbone is trained on random crops of 64 frames which corresponds to roughly 4 seconds. Additionally, RGB values were normalized and random color jittering was applied to each input channel. Video clips were randomly flipped horizontally with a probability of 50%. In the event a video was flipped, labels identifying bilateral states (such as "left" and "right") were fixed accordingly.

Table 5: Exercise options for the warm-up, main, and cool-down sections of the workout.

| Section | Exercise Candidates |
|---|---|
| Warm-up | jumping jacks, high knees, butt kickers, air jump rope, good mornings |
| Main | push-ups, plank taps, moving plank, squats, walking lunges, lunge jumps, puddle jumps, mountain climbers, floor touches, quick feet, squat jumps, squat kicks, standing kicks, boxing squat punches |
| Cool-down | deltoid stretch, quad stretch, shoulder gators, toe touchers |

https://doi.org/10.52202/079017-2415

### D.2 Short-clip Videos

We fine-tune the STREAM-VLM model end-to-end after initializing the 3D CNN vision backbone from the previous stage. The LM backbone is initialized with a pre-trained LLaMA-2-7B. We train the model for 2 epochs using a learning rate of $5 \times 10^{-6}$. We use a batch size of 32, gradient norm clipping of 1.0, and the AdamW optimizer [41] with betas $0.9$ and $0.95$ with a weight decay of $0.01$. The loss objective is a standard cross-entropy loss for next-token prediction.

**Data Preparation.** In this stage, we fine-tune using the fitness feedbacks and fitness questions annotations on the QEVD-FIT-300K collection. Sequences are prefixed with the system message below, followed by a sequence of <next> tokens equal to the video features length, and finally the question-answer pair from the dataset. The loss on all but the answer tokens and <next> are $0$ in this stage.

> <system>You are an expert fitness coaching AI who coaches users as they exercise. You observe them silently, assess their performance, and answer any questions they have.</system>

For short video clips, we always use the request "Please provide a feedback for the user." in place of the question.

### D.3 Long-range Videos

LoRA [28] is used for fine-tuning our model in this stage on QEVD-FIT-COACH with a learning rate of $1 \times 10^{-6}$ and LoRA dimension as 32. Other training details remain the same as the previous stage.

**Data preparation.** Due to memory constraints, we train on individual exercise segments which corresponds to a length of roughly 30 seconds. We leave it to future work to train across timed exercise transitions. Sequences are prepared as an interleaved sequence of <next> tokens and feedback response tokens. Feedback response tokens are prefixed with the special <feedback> action token and added in the sequence according to its ground truth timestamp. Finally, the sequence is prefixed with the following system prompt:

> <system>You are an expert fitness coaching AI who coaches users as they exercise. You assess their performance, and proactively provide feedback.</system>

## E Baselines and Evaluation

### E.1 LLM-Accuracy Prompt

We used the following prompt to evaluate the generated feedback:

> <INST> You are an intelligent chatbot designed for evaluating feedback sequences provided by a virtual fitness coach to a person. You always provide your responses as a python dictionary string.
>
> Your task is to compare the accuracy of the the predicted feedback with the ground truth feedback. Here is how you can accomplish this:
>
> -The predicted feedback must be factually accurate, relevant and align with the ground truth feedback.
>
> -Consider synonyms or paraphrases as valid matches.
>
> -Take into account repetition counts that can expressed both in numeric form or in words.
>
> Please evaluate the following predicted feedback:
>
> -Ground truth feedback: · · ·
>
> -Predicted feedback: · · ·
>
> Provide your evaluation as a python dictionary string with the accuracy score where the score is an integer value between 1 and 5, with 5 indicating the highest level of accuracy. Generate the response only in the form of a Python dictionary string with keys 'score', where its value is the accuracy score in INTEGER, not STRING. DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. For example, your response should look like this: "score": 3.2. </INST>

### E.2 Language-only Socratic Baseline

The following prompt was used to generate feedback with the socratic baseline shown in Tab. 4:


https://doi.org/10.52202/079017-2415

You are an expert fitness coaching AI who coaches users as they exercise. You assess their performance, count repetitions, and proactively provide feedback. The user should be doing · · ·

Provide a SHORT ONE SENTENCE RESPONSE to the user based on based on the activity from the last 5 seconds shown below. Take into account what you said before but DO NOT repeat it exactly. The response should be in second-person perspective. Ask them to correct any mistakes you see otherwise provide encouraging feedback.

For example:
User activity:
Timestep: 0.55 – The user is doing the exercise.
Timestep: 1.55 – The user is doing the exercise. Timestep: 2.45 – The user was fast.
Timestep: 3.45 – The user was fast.
Timestep: 4.62 – The user was fast.
Timestep: 5.42 – The user has good form.

Response: Woah, great speed there!

Previous user activity:
· · ·

This is what you've said so far for the previous activity:
· · ·

Latest user activity:
· · ·

Response: · · ·

Time-stamped user activity for the most recent 5 second window are included following "Latest user activity:" using the format shown in the example. The full history of previous activity and generated feedbacks are also provided in the prompt.

### E.3 Vison-language Baselines

The following prompt is used to generate feedback with the vision-language baseline models presented in Tab. 3.

You are an expert fitness coaching AI who coaches users as they exercise. You assess their performance, count repetitions, and proactively provide feedback. The user should be doing · · ·

This is what you've said so far over the last X seconds:
· · ·

Provide a SHORT ONE SENTENCE RESPONSE to the user based on what you see in the video. Take into account what you said before but DO NOT repeat it exactly. The response should be in second-person perspective. Ask them to correct any mistakes you see otherwise provide encouraging feedback.

**Number of frames.** The number of frames provided to the model vary across the baselines. For InstructBLIP, the latest frame is shown. For Video-LLaVA and Video-LLaMA, 8 uniformly sampled frames from the most recent 5 second window were shown. For Video-ChatGPT, LLaVA-NeXT-Video, and LLaVA-Vid, 20, 8, and 20, frames per 5 second interval were provided, respectively – frames were accumulated during evaluation for each feedback interval (*E.g.*, when generating a feedback at 10s, the frames from the first 5s interval are kept). The history of responses is provided in the prompt for all models.

### E.4 Alternative Evaluations

To ensure the accuracy of the LLM-based auto-evaluation method described in Sec. 5.1, we perform the evaluation with alternative LLMs and compare them to human evaluations. Human evaluations are done on a random subset of 200 feedbacks. As shown in Tab. 6, the overall ranking order is preserved across all evaluation methods.

Table 6: Evaluation of models fine-tuned with the FIT-COACH dataset on the FIT-COACH benchmark. ([†]indicates results of non-interactive models evaluated at regular intervals, [*]indicates human evaluation is conducted on a smaller set of 200 feedbacks.)

| Evaluation Method | Socratic-LLaMA-2-7B[†] | Video-ChatGPT [43] (fine-tuned)[†] | STREAM-VLM | STREAM-VLM (w/o 3D CNN) | STREAM-VLM (w/o Action-Tokens)[†] |
|---|---|---|---|---|---|
| Human[*] | 2.63 | 2.59 | **2.80** | 2.51 | 2.71 |
| Mixtral-Instruct-0.1 [29] | 2.39 | 2.42 | **2.56** | 2.17 | 2.56 |
| LLaMA-3-8B-Instruct [3] | 1.74 | 1.82 | **1.90** | 1.62 | 1.89 |
| LLaMA-3-70B-Instruct [3] | 2.17 | 2.33 | **2.45** | 2.11 | 2.41 |


https://doi.org/10.52202/079017-2415

# QEVD Datasheet [†]

## Motivation

**For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

The dataset and benchmark were created to enable automated live fitness coaching. It tries to fill two gaps with existing datasets: 1. **Fitness domain knowledge:** There are no publicly available datasets that provide examples of common mistakes, variations in form and intensity over a wide variety of datasets, 2. **Corrective feedbacks:** There are no publicly available datasets that provide long-range sequences of user with varied skill levels performing workouts along with corrective feedbacks to help the users successfully complete the workout in case of mistakes.

**Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

The dataset was created by the authors of the paper on behalf of Qualcomm Technologies Inc. and TwentyBN GmbH.

**Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

N/A.

**Any other comments?** None.

## Composition

**What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

The dataset consists of two types of video instances: short-clips and long-range videos as described in the main paper. The short-clips each cover one of 148 different exercises and their variations including: varied pacing, performing common mistakes, and modified form. The long-range videos are of participants performing a structured workout consisting of multiple exercises (4-6) while receiving live feedback.

**How many instances are there in total (of each type, if appropriate)?**

In total the dataset consists of $298,089$ and 223 instances of short-clip and long-range videos respectively.

**Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

Yes, the dataset contains all video instances that were collected by us.

**What data does each instance consist of? "Raw" data (e.g., unprocessed text or images) or features?** In either case, please provide a description.

Each instance consists of a video.

[†]Based on Gebru *et. al.* [21]

**Is there a label or target associated with each instance?** If so, please provide a description.

Each short form video has one or more of the following labels:

1. Fine-grained labels: The label specifies one of 148 different exercises and their variations including: varied pacing, performing common mistakes, and modified form. The full list of classes is provided with the supplemental material.
2. Fitness questions: This includes question-answer pairs that can be classified into - high-level, focusing on the general exercise type and performance, or fine-grained, targeting specific details of the exercise execution.
3. Fitness feedbacks: These feedbacks occur at the end of each video from a second-person perspective to support feedback in a live coaching session.

The long-range videos are annotated with feedbacks from a second-person perspective to support feedback in a live coaching session, along the timestamps of each recorded feedback.

**Is any information missing from individual instances?** If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

None.

**Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** If so, please describe how these relationships are made explicit.

N/A

**Are there recommended data splits (e.g., training, development/validation, testing)?** If so, please provide a description of these splits, explaining the rationale behind them.

Yes, the primary test set is the FIT-COACH benchmark, consisting of long-range videos with 7 participants. The train/validation set is the FIT-COACH dataset consisting of: long-range videos with 21 unique participants and short-form videos annotated with fine-grained labels, fitness questions and feedbacks.

**Are there any errors, sources of noise, or redundancies in the dataset?** If so, please provide a description.

We cannot fully exclude the possibility of human (annotator) errors in the labels/questions/feedbacks.

**Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

The dataset is fully self-contained.

**Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals non-public communications)?** If so, please provide a description.

No.

**Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** If so, please describe why.


https://doi.org/10.52202/079017-2415

No.

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

Yes.

**Does the dataset identify any subpopulations (e.g., by age, gender)?** If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No.

**Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** If so, please describe how.

While the faces of the individuals in the video are visible, the videos were collected under a direct agreement with the crowd workers, permitting research and commercial use. Audio and meta-data information from the videos were removed.

**Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** If so, please provide a description.

No.

**Any other comments?** No.

## Collection Process

**How was the data associated with each instance acquired?** Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The data was directly observable.

**What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** How were these mechanisms or procedures validated?

A simple web interface was used for recording videos and creating annotations. The resulting data was manually inspected to ensure data integrity.

**If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

N/A

**Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

The dataset was collected with the help of crowdworkers and contractors. All participants received appropriate and fair compensation for the regions where the participants were located.

**Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news**

**articles)?** If not, please describe the timeframe in which the data associated with the instances was created.

The recordings and annotations were created over the course of several months.

**Were any ethical review processes conducted (e.g., by an institutional review board)?** If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Manual review of a representative subset of the data was performed; audio an meta-data were removed.

**Does the dataset relate to people?** If not, you may skip the remaining questions in this section.

Yes.

**Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

The data was collected directly.

**Were the individuals in question notified about the data collection?** If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Yes.

**Did the individuals in question consent to the collection and use of their data?** If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

Yes.

**If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

Participants may reach us via email at `research.datasets@qti.qualcomm.com`.

**Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.

N/A

**Any other comments?**

No.

## Preprocessing/cleaning/labeling

**Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** If so, please provide a description. If not, you may skip the remainder of the questions in this section.

Continuous-valued labels, specifically those pertaining to exercise-specific speed and range of motion, were discretized.


https://doi.org/10.52202/079017-2415

**Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** If so, please provide a link or other access point to the "raw" data.

No.

**Is the software used to preprocess/clean/label the instances available?** If so, please provide a link or other access point.

No.

**Any other comments?**

No.

## Uses

**Has the dataset been used for any tasks already?** If so, please provide a description.

State of the art video-language models have been trained and tested on the data.

**Is there a repository that links to any or all papers or systems that use the dataset?** If so, please provide a link or other access point.

No.

**What (other) tasks could the dataset be used for?**

N/A

**Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?

The main aim of the dataset is to serve as a starting point for the development of models that can provide interactive feedbacks for fitness coaching. As such, predictions of models trained on the data cannot be guaranteed to be free from any bias. In addition, language models generally can produce harmful and biased content, make incorrect claims and produce wrongful advice. This needs to be taken into account when interacting with, deploying or building on these models, particularly in sensitive domains like fitness coaching where incorrect advice may lead to physical harm.

**Are there tasks for which the dataset should not be used?** If so, please provide a description.

N/A

**Any other comments?**

No.

## Distribution

**Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** If so, please provide a description.

Yes, the dataset is publicly available.

**How will the dataset will be distributed (e.g., tarball on website, API, GitHub)** Does the dataset have a digital object identifier (DOI)?

The dataset can be downloaded from: `https://www.qualcomm.com/developer/software/qevd-dataset` as a ZIP archive.

**When will the dataset be distributed?**

The dataset is publicly available currently.

**Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.

The dataset is released under a CC BY-NC-ND 4.0 license (`https://creativecommons.org/licenses/by-nc-nd/4.0/deed.en`).

**Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.

No.

**Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.

No.

**Any other comments?**

No.

## Maintenance

**Who will be supporting/hosting/maintaining the dataset?**

The dataset is hosted and maintained by Qualcomm Technologies Inc.

**How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

`research.datasets@qti.qualcomm.com`

**Is there an erratum?** If so, please provide a link or other access point.

N/A

**Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?

Yes, updates will be communicated on the website `https://www.qualcomm.com/developer/software/qevd-dataset`.

**If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** If so, please describe these limits and explain how they will be enforced.

No.


https://doi.org/10.52202/079017-2415

**Will older versions of the dataset continue to be supported/hosted/maintained?** If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

Yes. All versions should be available at `https://www.qualcomm.com/developer/software/qevd-dataset`.

**If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.

No.

**Any other comments?**

No.