

ExpertAF: Expert Actionable Feedback from Video

Supplementary Material

A. Supplementary video

We provide a supplementary video containing an overview of the paper. The video contains details of the data collection approach, the method idea and finally, qualitative and quantitative results, **with video examples**.

B. Expert feedback dataset

In this work, we propose a novel expert feedback dataset that contains the tuple $(\mathcal{V}, \bar{\mathcal{V}}, T)$. The following subsections contain additional details about the specific prompt used, some limitations of the dataset and finally, some visualizations. Please also refer to the supplementary video for video demo of the dataset samples.

B.1. Prompt for expert commentary classification and body region tagging

In Sec. 3.2, we outline the process of obtaining a weakly-supervised training data from commentaries in Ego-Exo4D [29]. The first step involves summarizing the expert commentary into a single actionable feedback, followed by tagging the body region to whether the execution is correct, incorrect or not mentioned at all. We use the following prompt to ask Llama 3-70B [2]:

System: Llama3 default system prompt.

User: The following is an expert commentary about a person playing Basketball. Give a one sentence summary of the expert feedback and then mention which body parts out of Head, Shoulder, Hands, Arms, Legs, Jump needs improvement and which ones are good execution (you can choose multiple body parts). Here is the expert’s commentary:

He came down on one foot that time. You want to make sure you come down on two feet. Right now he’s putting a little bit too much pressure and stress. Then on top of that, his left knee is locked, which could easily cause some hypertension. So be very aware and careful of your landing.

Assistant: One sentence summary: He came down on one foot and his left knee is locked, which could cause some hypertension. \n Needs improvement parts: Legs, Jump. \n Good execution parts: None.

User: Let’s take a look at the placement of the shooter’s left guy hand. You can see that it’s in a really good position on the left side of the ball. If you take a look at the shooting hand, his right hand is underneath the ball with his right index finger in

the middle section of the ball. His right index finger is in the middle section of the ball. This is good positioning of both his right shooting hand and his left guy hand on the ball.

Assistant: One sentence summary: Shooter’s hand is in a really good position on the left side of the ball. \n Needs improvement parts: None. \n Good execution parts: Hands.

User: <new_question>

Assistant:

This example prompt is the template for basketball, and we follow a similar process for other scenarios. To ease the generation process, we provide one incorrect and one correct execution example, shown above. Subsequently, we parse the outputs and discard the commentaries that cannot be parsed. Most of the discarded commentary contains no information to reliably classify it, as done above. As included in the prompt above, we use the following body regions for all the scenarios for a coarse classification:

Head, Shoulder, Hands, Arms, Legs, Jump

B.2. Expert feedback classification examples

Fig. 5 shows some examples of expert commentary preprocessing. The above described prompt results in these samples. We obtain a summarized expert commentary, along with label of correct and incorrect execution for every body region. Recall that the expert commentaries are obtained by converting expert speech to text using ASR. Consequently, some samples contain incomplete information and, hence, discarded (see bottom right in Fig. 5).

B.3. Visualization of the dataset

Fig. 6 shows six samples from the weakly-supervised expert feedback dataset. See the video examples in the attached supplementary video. Each example contains the tuple $(\mathcal{V}, T, \bar{\mathcal{V}})$. All the learner demonstrations have errors that are pointed out in the expert commentary (center). Finally, the error is corrected in the expert demonstration. For example, the jump is incorrect (Example 1) or the follow-through after shooting is incorrect (Example 6). The same is corrected in the expert demonstration. Note that all the videos in the dataset are accompanied by 3D poses.

B.4. Dataset statistics

As noted in Sec. 3.4, we use Ego-Exo4D [29] for our experiments. We choose the physical scenarios in the dataset—

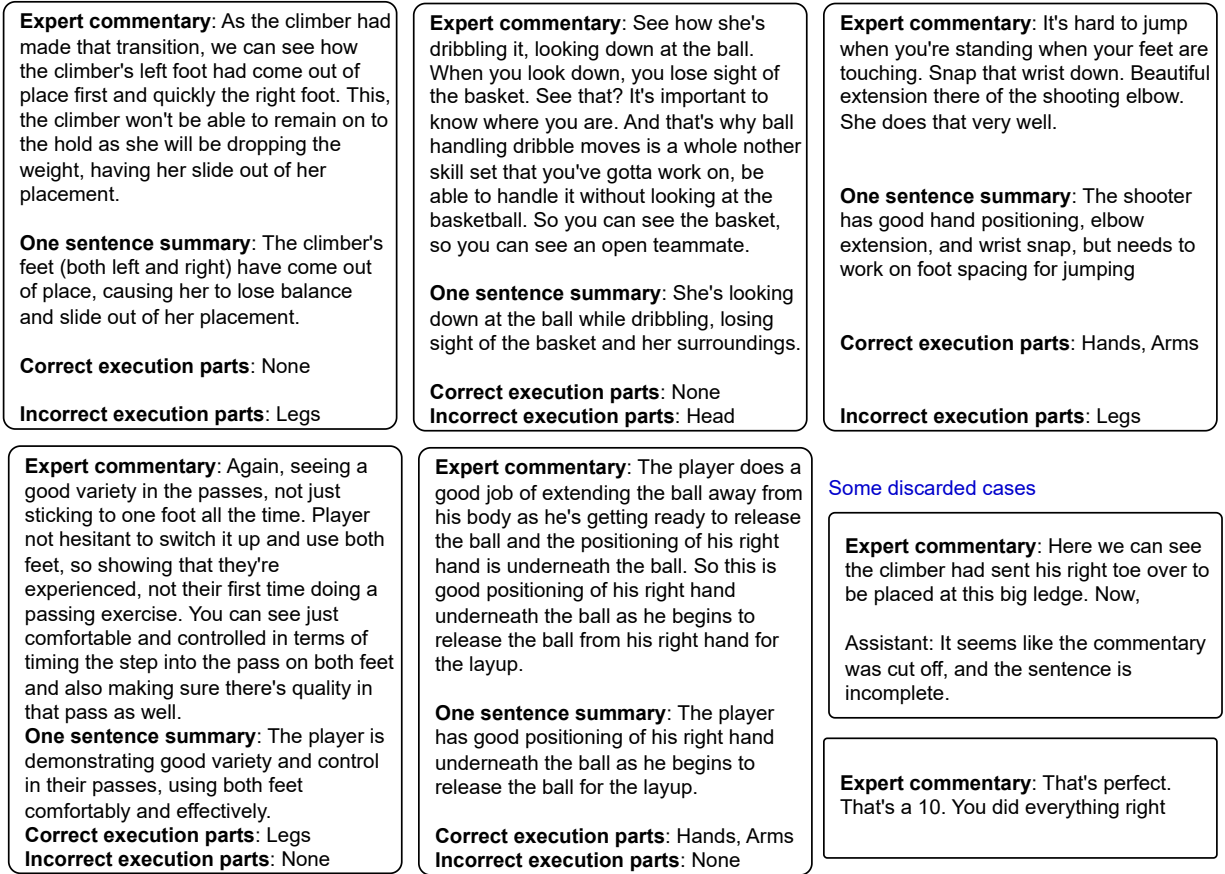


Figure 5. **Expert commentary classification examples.** Llama3 correctly generates a one-sentence summary of the expert commentary, along with a tagging of the body region with incorrect and correct execution. The last example shows two discarded examples.

basketball, rock climbing and soccer. These physical scenarios contain 34092 expert commentaries. The expert commentary classification and body region localization results in labeling 16791 commentaries as having incorrect executions of at least one body region and 20946 commentaries with at least one good execution. Some commentaries do not contain enough information for classification and are discarded. There are 2073 distinct participants in the scenarios of interest and those are classified into four categories—novice (340), early expert (420), intermediate expert (642) and late expert (671). The created expert feedback dataset contains 25505 training and 1272 testing samples. Each training and testing clip is 4 seconds long, which is the typical duration of an execution in the selected physical scenarios.

B.5. Manual verification of the test set

We verify the automatically curated test set to obtain a clean subset of 1,272 samples. The verification process firsts

check if the expert commentary is correctly summarized into a single sentence, along with a correct assignment of the body region. Next, we examine the tuple $(\mathcal{V}, T, \bar{\mathcal{V}})$ for correctness. We check if the feedback in T is incorporated in $\bar{\mathcal{V}}$. We only keep the segment if we are certain that the expert demonstration corrects the mistake in T . Overall, only 5.5% of the samples were discarded—showcasing the quality of the automated pipeline.

C. Additional implementation details

The demonstration clips are 4 seconds long and we sample frames at 32 fps. The InternVideo2 [77] video encoder f_V takes in 8 frames and thus, generates 4 features per second (totalling 16 features). We concatenate the ego and exo features to create 32 input tokens for each demonstration. In the expert demonstration retrieval training, we use the log likelihood loss as the relevance score, lower is better.

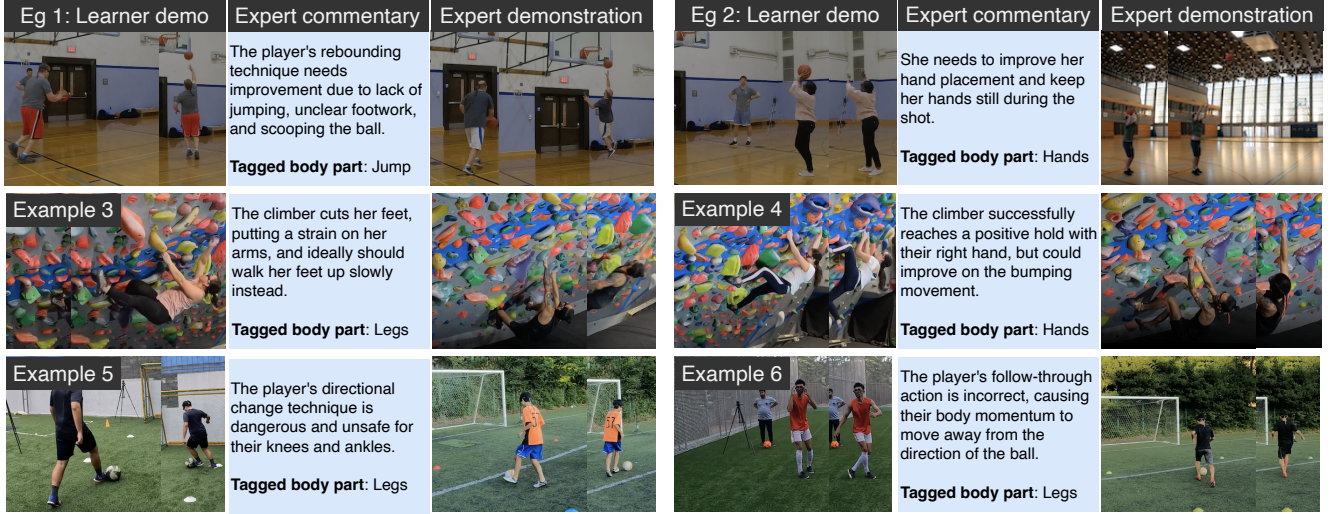


Figure 6. **Examples from the expert feedback dataset.** Some examples of the expert feedback tuples $(\mathcal{V}, T, \bar{\mathcal{V}})$ generated by our approach. Notice how the expert demonstration corrects the errors in the learner demonstration. For example, the top left video shows the person attempting a shot without jumping, as noted in the expert commentary. This error is corrected in the expert demonstration, where the person jumps correctly when shooting.

D. Additional ablations

In addition to the ablations discussed in Sec. 4, we also evaluate the choice of the LLM \mathcal{L}_s , the contribution of ego and exo videos, and the joint training with all scenarios. Table 2 summarizes the performance for all these ablations. We discuss each of them below:

Effect of the choice of the LLM. We observe that the performance increases with the relative strength of the LLM. This experiment suggests that the expert actionable feedback will further improve with advancements in these strong language models. Notably, our method will still be useful to learn the fine-grained differences between learner and expert demonstrations, and to provide actionable expert feedback.

Effect of separate training. We jointly train all the three scenarios—*basketball*, *soccer* and *rock climbing*. We observe that training all scenarios separately does not improve the performance. The performance with separate training remains lower due to no cross-scenario learning and lower generalizability.

Contribution of ego and exo views. We observe a better performance for exocentric-only ablation. This result is expected since exo view captures the body pose more accurately. However, if we use pose + egocentric views (‘w/ ego + pose’), the performance increases since the body pose compensates for the missing exocentric video.

Generation vs retrieval when the expert demonstration is missing in the candidate set. To evaluate the quality of the generation in the absence of correct demonstration in the retrieval candidate set, we remove the 50 closest sam-

ExpertAF	Comm. Gen.			D.R.		P.
	B@4	M	RL	R	mR	P ↓
w/ Llama3.2 1B	42.8	48.5	53.0	17.4	177	153
w/ Llama3.2 3B	43.6	49.1	53.8	18.3	168	145
w/ sep. train	43.2	48.0	52.6	15.6	176	155
ret. w/o GT	—	—	—	—	—	160
w/ ego only	40.0	45.7	50.1	15.8	194	—
w/ exo only	44.2	49.0	54.1	18.3	169	—
w/ ego + pose	44.4	49.2	54.0	18.3	164	144
Ours	44.9	49.6	54.6	19.1	158	135

Table 2. **Results of additional ablations.** Our method outperforms all ablations. See the text for discussion. (Comm. Gen.: Expert commentary generation, D.R.: Expert demonstration retrieval, P.: Expert pose generation).

ples with the ground truth from the retrieval set, to simulate missing correct demonstrations. Our retrieval method finds the best remaining demonstration, and extracts pose. See Tab. 2 “ret. w/o GT”. The PA-MPJPE error is 160 mm which is worse than the error in the third task.

E. Additional result visualization

Fig. 8 contains additional result visualization. We see that our method generates better expert commentary compared to all baselines. Similarly, our expert demonstration retrieval and pose generation correct the mistakes in the input demonstration. **See the video visualization in the attached supplementary video.**

Expert commentary human evaluation form

Instructions

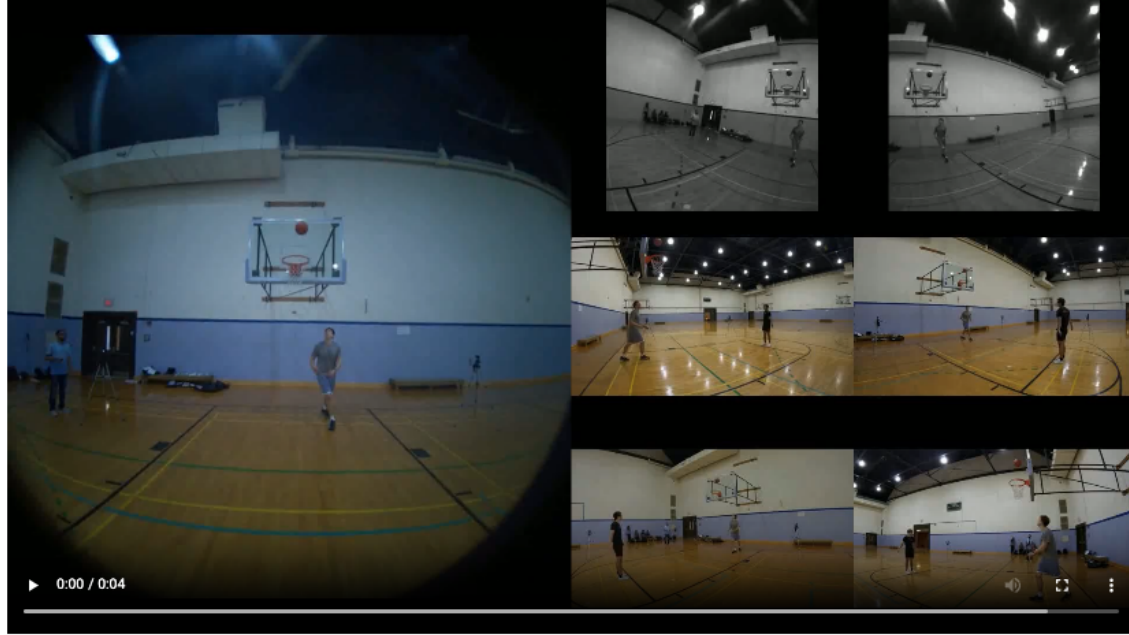
Please watch the video and the accompanying expert commentary generated by a machine learning model. Based on your knowledge of the sport, determine whether the expert commentary is correctly generated or not. The options are correct and incorrect with two confidence choices.

After every 10 samples, we have added a 'Save Progress' button that saves the progress in the cache so that it can be recovered in case the browser is closed. When you are done, press 'Download Progress' and send it to me.

Name:

Email:

Video 1:



Generated commentary:The player needs to improve his follow-through and release point to achieve a better arc on his shot.

Choose one of the following options:

- ☐ Correct commentary, confident
- ☐ Correct commentary, not confident
- ☐ Incorrect commentary, not confident
- ☐ Incorrect commentary, confident

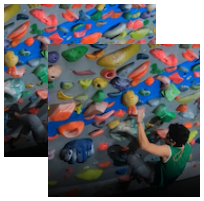
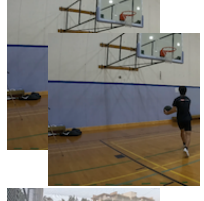

Figure 7. **Human evaluation interface.** A screenshot of the human evaluation interface.

F. Limitations

We observe the following limitations:

Incorrect 3D pose for some samples. Recall that the 3D pose sequence in [29] is calculated by triangulating the position from various exo cameras placed around the subject. Thus, the 3D pose sequence is an *auto-ground truth*. Therefore, some samples have reconstruction error. However, this issue is beyond the scope of our work.

Variable expert commentary granularity. There are many experts annotating the dataset. Therefore, some experts give fine-grained feedback about legs, hands etc. whereas some experts provide coarse instructions like incorrect posture. A consistent granularity is desirable for a better training, but we do not ignore any sample for a more diverse expert commentary generation.

	Ground truth The climber is wavering and fatigued, struggling to get their feet up and hold the lock off.	Ours The climber is struggling to maintain the position and is losing tension in their body.	LLaVA-FT The climber needs to jump closer to the wall to ensure they don't fall off.	VideoChat2 One way they can improve is by incorporating different exercises to challenge their muscles.	PoseScript The person is in a sitting position with their legs slightly bent and the right arm raised and bent at the elbow. The left arm is bent at the elbow.
	The player's dribbling and stepping technique needs adjustment to avoid traveling and create spacing.	The player's layup attempt was unsuccessful due to poor footwork and lack of control.	The player should keep the head towards the basket for a better knowledge of the surroundings.	The person can improve their physical activity by incorporating more variety.	The person is in a dancing pose with their legs crossed over their right leg. Their head is tilted back and to the left. Both arms are extended behind them.
	The player's kicking technique lacks control and power due to incorrect foot positioning and follow-through.	The player is not able to generate power due to lack of correct foot position.	The player is moving fast when dribbling and needs to slow down.	It's essential to have a good base level of fitness and strength to make the experience safer and more comfortable.	The person is lying down on their right side with their left leg extended back and their left foot is on the ground. Their right arm is bent at the elbow.



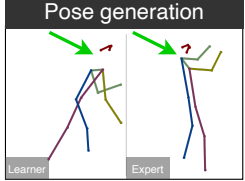





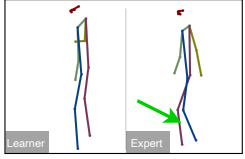


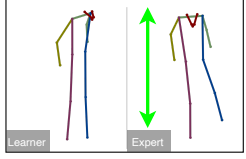
Learner demonstration 	Expert comment. - Ours The player needs to change their head position to look at the basket when dribbling.	Ground truth commentary The player needs to keep their eyes on the target and maintain elevated arms while dribbling.	Expert demo. retrieval 	Pose generation 
	The player is not jumping when shooting the ball.	The player needs to jump on their shot to give the ball more arc and momentum to get above the rim.		
	The player is letting the ball go far from the body resulting in a disbalance.	The player's failure to move their body in sync with the ball is causing them to make off-balance contact with the ball		
	The player's body positioning and footwork are affecting their ability to control the ball.	The player needs improvement in foot positioning and body alignment for better control and balance.		

Figure 8. **Additional results visualization.** (Top) Comparison of expert commentary generated by various baselines. (Bottom) Examples of expert commentary generation, demonstration retrieval, and pose generation by our method.

G. Human evaluation details

G.1. Evaluation setup

We obtain a random subset of 250 samples (out of 1272) for human evaluation. For each scenario, we ask 5 raters to evaluate all the methods. As mentioned in Sec. 4, we group the methods for efficient rating, scoring 5 groups of baselines for each of the three scenarios. The raters are appropriately compensated. To reiterate, we ensure all the raters

for each scenario have 2+ years of experience of doing that physical scenario.

G.2. Evaluation interface and instructions

Fig. 7 shows the human evaluation interface, along with the instructions. We have similar interface for expert commentary retrieval, and pose generation. The sampled are randomized for each method to avoid bias while annotating.