

## A Supplementary material for Learning Skill-Attributes for Transferable Assessment in Video

### A.1 Table of content

This supplementary contains the following:

- **A supplementary video** that first motivates the problem with some example actionable feedback requests from learners on Reddit (discussed in Sec. 3.3). Next, we show qualitative results from Ego-Exo4D [36], QEVD [67], and YouTube videos. Finally, we also show some failure cases.
- **Sec. A.2: LLM prompt** that we use to obtain skill-commentary from expert commentary.
- **Sec. A.3: Additional quantitative** results to show metrics for skill-attribute generation, and QEVD [67] zero-shot results.
- We discuss **limitations** in **Sec. A.4**.
- We also discuss **societal impact** in **Sec. A.5**.

### A.2 LLM prompt for obtaining skill-attribute from expert commentary

In Sec. 3.3, we discuss using a large language model (LLM) to extract skill-attributes that are suboptimally performed. We use the following prompt:

**System:** Answer the question regarding a commentary about a sports drill. Do not add information not present in the question.

**User:** The transcript of an expert commenting on a SPORT NAME COMES HERE drill is given below. List down the concepts that are correct and incorrect in the drill, as noted by the expert. The concepts are distinct aspects of the skill, e.g., control, body positioning, speed, body movement, hand position, and so on. Feel free to come up with newer concepts and write the response in two lines. The first line should contain the correctly shown concepts, and the second line should contain the incorrectly shown concepts. It should be in this format. Correct - comma separated concepts.  
Incorrect - comma separated concepts.

Here is the expert feedback:

NARRATION COMES HERE

**Assistant:**

We prompt the LLM to provide both correctly and incorrectly demonstrated skill-attributes. Our Attribute-Retrieval baseline (Sec. 4.3) uses both of them.

### A.3 Additional quantitative results

In Sec. 4.2 and Tab. 1 (top left), we show results for generating skill-attributes for IoU@0.7. We extend the table to show results on IoU@ $k$  for  $k \in \{0.7, 0.8, 1.0\}$  in Tab. 3.

Next, we show zero-shot transfer performance on the QEVD [67] dataset (summarized in main paper, and detailed here due to space limitations). Note that all videos in QEVD are for fitness exercises, and they are not as distinct as a different sport. Moreover, every video contains multiple exercises, with the transition labeled as “Moving to (EXERCISE NAME)...”. We use these labels to split the videos per-exercise. We discard instances that are before the start of any labeled exercise. Next, we create sport labels based on similarity in execution and effects. We create this division for the purpose of zero-shot transfer experiments. This split is created using consensus from ChatGPT-4o [2] and Llama-3.1 [4], and finally manually verified. The splits and the reasonings are given in Tab. 2. A total of 23 unique exercises are divided into 5 groups.

Fig. 6 shows the results. First, in skill-attribute generation (top row), we see that our method outperforms both Stream-VLM [67] and InternVideo2 [87] baselines. Moreover, as seen in Ego-Exo4D [36], the performance decrease in the zero-shot setting is milder than the drop observed in the

Table 2: Buckets of exercises in QEVD [67] grouped by similarity in execution and effect.

Group name	Exercises	Remark
Stretches & mobility	quad stretch, armcrosschest, good morning beginner, floor touches, toe touchers	Focused on flexibility and range of motion; often used in warm-up or cooldown phases. Involves static or slow dynamic movement.
Cardio & agility	high knees, quick feet, jumping jacks, air jump rope, butt kickers, puddle jumps	Elevates heart rate with low to moderate resistance; emphasizes agility and coordination with repetitive footwork.
Leg strength & lower-body	squats, squat jumps, squat kicks, walking lunges, lunge jumps, standing kicks	Targets glutes, quads, hamstrings through controlled or explosive leg movements. Builds strength and endurance.
Core & upper-body	plank taps, moving plank, pushups, shoulder gators	Focuses on core stabilization and upper body strength, particularly arms, shoulders, and chest. Often bodyweight-based.
Full-body	boxing squat punches, mountain climbers	High-intensity, compound movements that engage multiple muscle groups while promoting coordination and rhythm.

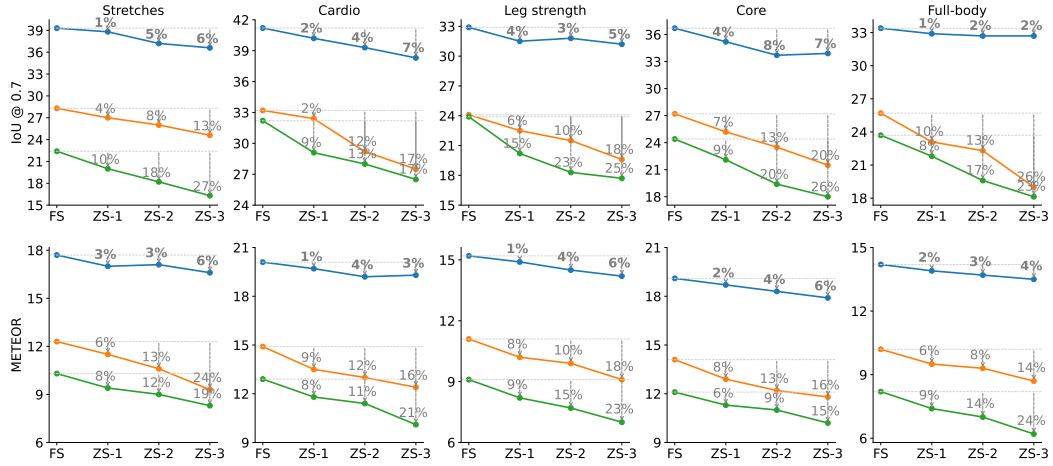


Figure 6: **Zero-shot performance.** Performance trend when testing on various skills (stretches, cardio, etc.) for different in-domain and zero-shot training settings (FS, ZS-1, etc.) for skill-attribute generation (top) and actionable feedback generation (bottom) for QEVD [67]. Legend: — CROSSTRAINER, — Stream-VLM [67] and — InternVideo2 [87].

baselines. Finally, we see a similar trend in actionable feedback generation (bottom row). Overall, zero-shot results in both Ego-Exo4D [36] and QEVD [67] show that our idea of learning to generalize using skill-attributes is effective.

#### A.4 Limitations

We observe that the proposed model struggles with feedback that is about aspects not directly visible in the video. Phrases like *lacking intent* is not groundable definitively and hence, is not captured. Nevertheless, this inability to capture abstract notions is also observable in all the baselines, and in general, vision encoders. Secondly, as we discuss in Sec. 4.1, commentary is subjective and there can be various correct ways of providing feedback that improves a learner’s performance.

Table 3: **Quantitative results.** Skill-attribute generation results for IoU@ $k$  for  $k \in \{0.7, 0.8, 1.0\}$  for Ego-Exo4D [36] and QEVD [67], extension of Tab. 1 (top left) on remaining  $k$  values.

Method	IoU	@0.7	@0.8	@1.0	Method	IoU	@0.7	@0.8	@1.0
InternVideo2-NN [87]	14.0	8.9	7.7		InternVideo2-NN [87]	23.8	16.3	14.8	
InternVideo2-FT [87]	15.0	9.4	8.2		InternVideo2-FT [87]	24.5	16.6	15.3	
VideoChat2 [46]	9.3	6.6	4.0		VideoChat2 [46]	16.9	11.4	10.1	
LLaVA [52]	9.7	7.2	4.9		LLaVA [52]	17.3	12.5	11.8	
LLaVA-FT [52]	14.6	9.1	8.1		LLaVA-FT [52]	26.9	19.2	18.2	
Stream-VLM [67]	14.5	9.1	8.3		Stream-VLM [67]	28.0	19.9	18.6	
ExpertAF [8]	15.0	9.5	8.4		ExpertAF [8]	28.1	19.7	18.3	
Attribute-Retrieval	19.7	12.7	10.7		Attribute-Retrieval	32.4	24.7	23.0	
<b>CROSSTRAINER</b>	<b>25.7</b>	<b>15.9</b>	<b>14.4</b>		<b>CROSSTRAINER</b>	<b>37.6</b>	<b>29.8</b>	<b>28.1</b>	

## A.5 Societal impact

Our CROSSTRAINER can be used for learning skills, especially long-tailed low-resource sports like *kho-kho*, *shinty*. On the positive side, our model democratizes access to skill coaching, and it promotes inclusivity in underrepresented sports. More learning will promote more people playing the sport, and eventually, more data for training and expansion of knowledge. However, the model is trained with Ego-Exo4D [36] and QEVD [67] that might have regional bias. We believe as more data is available, the biases will go down, and we will move closer towards full physical skill understanding.