Learning Skill-Attributes for Transferable Assessment in Video

Kumar Ashutosh
Univeristy of Texas at Austin

Kristen Grauman
University of Texas at Austin

Abstract

Skill assessment from video entails rating the quality of a person's physical performance and explaining what could be done better. Today's models specialize for an individual sport, and suffer from the high cost and scarcity of expert-level supervision across the long tail of sports. Towards closing that gap, we explore transferable video representations for skill assessment. Our CROSSTRAINER approach discovers *skill-attributes*—such as balance, control, and hand positioning—whose meaning transcends the boundaries of any given sport, then trains a multimodal language model to generate actionable feedback for a novel video, e.g., *"lift hands more to generate more power"* as well as its proficiency level, e.g., *early expert*. We validate the new model on multiple datasets for both cross-sport (transfer) and intra-sport (in-domain) settings, where it achieves gains up to 60% relative to the state of the art. By abstracting out the shared behaviors indicative of human skill, the proposed video representation generalizes substantially better than an array of existing techniques, enriching today's multimodal large language models. Project page: https://vision.cs.utexas.edu/projects/CrossTrainer/.

1 Introduction

The basis of assessing skilled physical activities, particularly sports, is largely visual. Precisely how a tennis player grasps and swings their racquet; how a basketball player releases the ball to shoot a free throw; how a rock climber stretches and pulls to traverse the boulder—such visual details are discernible to the expert eye and essential for providing meaningful coaching. Advances in multimodal video understanding could, therefore, transform AI-assisted coaching and skill assessment. For example, future AI agents could provide personalized feedback to users based on videos captured on their phone or smartglasses, greatly expanding the accessibility of 1-1 coaching. Similarly, AI could analyze how multiple players' skills would complement each other when building a team, or even detect patterns in injuries as a function of execution style.

Towards achieving the above, the core vision and machine learning task is as follows: given a video of an athlete's performance, estimate their skill level [12, 28, 37, 39, 67, 70, 71, 73] and indicate what could be improved [9, 16, 68].

This task raises important unsolved challenges. First of all, assessing skill requires *fine-grained* understanding of all physical aspects. Whereas traditional action understanding emphasizes high-level semantics [17, 18, 27, 55, 60, 87, 88, 92]—and thus seeks *invariance* to execution differences—here the subtle differences are exactly what matters. For example, a novice and expert shooting a soccer ball into the net might accomplish the same overall action, but details about their approach, footwork, and trajectory of the ball are essential to analyze skill. Secondly, compounding the challenge, supervision is costly and difficult to scale across the *long tail* of sports. An estimated 8,000 unique sports exist today across the world, but only a small subset is widely recorded and distributed—often driven by geographic and economic factors. Similarly, expert supervision is expensive and cannot be scaled up easily using traditional annotator pools. Making it worse, due to the common assumption that the axes of evaluation vary so wildly between sports, all prior work trains and tests on *in-domain* data, i.e., the same drill, exercise, or sport [9, 12, 37, 70, 70–72, 85, 98].

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-5359

Figure 1: **Overview of the idea.** Given a short video of an athletic skill, what could be better? Given video demonstrations from multiple sports, we learn skill-attributes that are incorrectly demonstrated, e.g., wrong foot positioning in badminton (top). These skill-attributes are common across various sports, and transfer to novel uncommon sports, e.g., shinty (bottom). Our method improves both the in-domain and zero-shot settings. Sports chosen for illustration; see Sec. 4 for dataset details.

We propose CROSSTRAINER, a new approach to video-based skill analysis that accounts for these challenges. Our key insight is to learn a universal *skill-attribute* representation that transcends the boundaries of sports and hence allows sharing and transfer between them. A skill-attribute is a describable fine-grained concept about the physical performance that can take on different visual forms in different sports, e.g., *ball control*, *foot positioning*, *coordination*, *timing*, or *balance*. By automatically learning these shared properties from video-language commentaries, and then surfacing them during training, we aim to amplify the value of the limited training data available for any one sport.

Building on the skill-attributes, we then train a generative multimodal language model to predict i) which skill-attributes are incorrectly demonstrated in a given video, ii) what the proficiency level is, and iii) free-form commentary about which adjustments would improve the performance. Essentially we factor skill assessment into its generalizable cross-sport component (e.g., *lacks control*) and its focused sports-specific component (e.g., *"you should increase the spacing between your legs for better control of the soccer ball"*). Our formulation aims to unlock the translation of skill assessment from widespread, highly marketed sports to uncommon, low-resource sports, such as *kabaddi*, *frisbee*, *waterpolo*, *shinty*, *kho-kho*, or *bandy*. See Fig. 1.

Though this represents a departure from today's AI models [9, 12, 37, 70, 70–72, 85, 98], cognitive science supports both transferability across sports having similar skill sets and shared terminology for assessment [19, 65, 77, 82]. For example, basketball players make better decisions playing soccer compared to tennis players, due to the former two sports' shared underlying properties [77]; similarly, contact-sport athletes exhibit better zero-shot understanding of a new contact sport compared to non-contact sport athletes [19]; and transferable skills arise from similar affordances [43, 81] or environmental properties [82]. Our work is the first to translate these intuitive findings from cognitive science into a working video understanding model.

We validate our ideas on three diverse datasets: Ego-Exo4D [37], which contains soccer, basketball, and rock climbing; QEVD [68], which contains fitness exercises; and in-the-wild YouTube videos of people tutoring physical activities. We show that training an intermediate skill representation yields both superior *in-domain* performance (for familiar sports) as well as better *zero-shot* performance (where the particular sport or drill is never seen by the model during training). CrossTrainer outperforms all competitive baselines on all axes of assessment—actionable feedback, skill-attributes, and proficiency estimation—with relative gains up to 60%. Furthermore, compared to any of the baselines, CrossTrainer shows much more graceful degradation when transferring to a novel sport. Overall our approach is a stepping stone for learning a unified skill-centric video representation for assessment and coaching in the presence of real-world data and annotation constraints.

2 Related Work

Learning representations from videos. Substantial research explores new ways to learn video representations [6, 48, 51, 74, 92], targeting video understanding tasks of action recognition [36, 45, 49, 89], action anticipation [1, 32, 33, 35, 57], procedural understanding [13, 14, 20, 62, 101], and temporal step localization [3, 58, 97, 99, 100]. Instructional videos offer a valuable window into skilled human activity [59, 84, 103], and recent work explores how to navigate between related how-to's [7, 8, 102] and identify their differences [61]. Whereas prior work explores activity-centric representations emphasizing the semantics of *what* is being done, our problem requires capturing *how* it is being done, which we show is essential for skill assessment.

Video-based sports analytics. Sports analytics and assessment raises a number of interesting challenges for computer vision [12, 37, 70, 71, 73, 85, 98]. Multiple ongoing workshops and challenges [22, 34, 64] offer tasks including ball spotting, foul recognition, and game state reconstruction. Prior work on skill assessment either assigns a score (or equivalence class label) to a video demonstration [12, 37, 67, 70, 71, 73], optimizes a group contrastive score distribution [85, 98], or chooses the better of two demonstrations [12, 29]. Broadening the scope of skill assessment beyond scoring videos, ExpertAF [9] aims to provide *actionable feedback* in the form of natural language commentary, relevant video retrievals, and generated pose corrections. As discussed above, all the above existing work in video-based skill assessment is limited to in-domain testing, whereas we explore transfer *across* sports, enabled by the proposed skill-attributes. In addition, orthogonal to the transfer contribution, our results across three datasets representing 6 distinct sports and fitness activities and 30 distinct drills raises the bar in breadth of validation compared to any prior skill assessment work.

Zero-shot generalization and attributes. Testing on novel classes and scenarios is crucial for real open-world settings. In one line of work, zero-shot generalization is enabled by shared multimodal representations learned from noisy vision-language data, benefiting image classification [75, 90, 93], action recognition [10, 95, 96], video to text and text to video retrieval [6, 80, 87, 88, 91, 92], or image segmentation [15, 26, 40]. In another line of work, *attributes* are intermediate variables [41, 42, 54, 66, 83, 94] that can express a new category even without training images (e.g., zebras are *black and white* and *striped*). Though they share our motivation for shared representations, all of the existing methods focus on semantics (what) as opposed to dynamic execution (how), and none are directly applicable to skill assessment from video. Ours is the first work to explore attributes for fine-grained activities in video and the first to demonstrate the relevance for skill feedback.

3 Method

We introduce the problem statement in Sec. 3.1, followed by an overview of key datasets (Sec. 3.2), our idea to extract skill-attributes (Sec. 3.3), the full model (Sec. 3.4), and training (Sec. 3.5).

https://doi.org/10.52202/085713-5359

Figure 2: **Discovered skill-attributes** from Ego-Exo4D [37] (left) and QEVD [68] (right). We see phrases reflecting generalizable physical concepts like control, hand/body positioning, and movement.

3.1 Problem definition

At inference time, given a short video clip $V \in \mathcal{D}_{te}$ from the test set $\mathcal{D}_{te}$, we want to assess its quality, even if the exact same skill/drill was never seen in training—and even (more extreme) if the sport in V was not seen during training. The desired assessment output covers three aspects: report the skill-attributes performed incorrectly, generate actionable feedback that can help the learner improve, and finally, estimate a proficiency score. Consistent with the transfer observed in cognitive science studies [43, 81, 82], we focus on physical skills executed by an individual and hence assume the video contains one person of interest; multi-player team interactions are also interesting but less amenable to transfer and outside the scope of this work.

To handle this task of assessing both in-domain and novel data, we employ a two-stage training process. In the pretraining stage, we train the model to generate the incorrectly demonstrated skill-attributes, i.e., we learn a function $\mathcal{F}_a$ to predict the skill-**a**ttributes $\hat{S} = \{s_1, s_2, ...\}$ that the person in the video should improve on:

$$\mathcal{F}_a(V \mid \mathcal{D}_{tr}) = \hat{S}, \tag{1}$$

where $\mathcal{D}_{tr}$ denotes the training set, e.g., if a person in the video is dribbling a soccer ball, S can be *control* and *leg positioning*, supposing the person is incorrectly executing those two. See Fig. 1, left.

In the second stage, we finetune the model for the remaining two aspects of assessment. Firstly, we generate feedback conditioned on the inferred skill-attibute set $\hat{S}$,

$$\mathcal{F}_t(V, \hat{S} \mid \mathcal{D}_{tr}) = T, \tag{2}$$

where T is a **t**extual actionable feedback statement for improvement, e.g., *"the player is too straight and should bend more for a better control"*. While the skill-attributes are words or short phrases indicating suboptimal dimensions in general terms, the feedback consists of sentence(s) elaborating on what to fix in the context of the observed sport. Finally, we generate the **p**roficiency level of the person in V, again conditioned on $\hat{S}$:

$$\mathcal{F}_p(V, \hat{S} \mid \mathcal{D}_{tr}) = P, \tag{3}$$

where P is the proficiency class label, e.g., *novice, intermediate, early expert* or *late expert*.

All aspects of assessment are evaluated in both the *in-domain* and *zero-shot* settings, to explore both the absolute performance of our proposed approach with respect to the literature (in-domain) as well as its ability to transfer to novel sports and scenarios (zero-shot). In the experiments we tackle multiple datasets of various sports and fitness activities, introduced next and detailed more in Sec. 3.5.

3.2 Skilled physical activity datasets

Before introducing our model, we next overview the key existing datasets leveraged in this study, since being familiar with their contents will help visualize our overall learning paradigm. **Ego-Exo4D** [37] contains 2,593 videos, totaling 239 hours with 289 total participants playing 3 sports—soccer, rock climbing, and basketball. **Qualcomm Exercise Videos Dataset (QEVD)** [68] has 223 long home fitness exercise videos, totaling 13 hours, where 28 total participants perform 23 structured workout

https://doi.org/10.52202/085713-5359

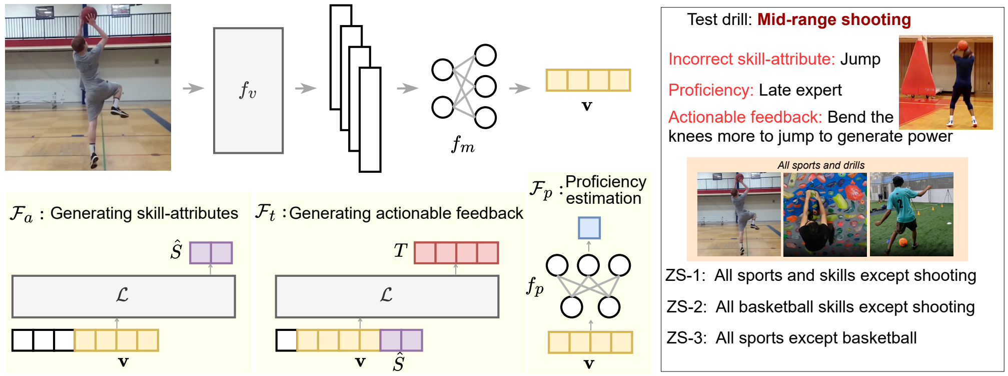

Figure 3: **Method overview and evaluation settings.** (Left) We encode videos into tokens $\mathbf{v}$ that can be fed to a multimodal LLM $\mathcal{L}$, with a mapper f_m that trains for skill-attributes. We use these visual tokens to generate skill-attributes (bottom left). Next, this pretraining is used to generate actionable feedback (bottom middle) and proficiency score (bottom right). (Right) Example of the various training settings for in-domain and zero-shot.

moves (like jumping jacks, planks, squats, etc.). To stretch our model on in-the-wild zero-shot settings, we also show qualitative results using **YouTube sports** videos collected by us—*frisbee*, *water polo* and *soccer (juggling)*.

Both Ego-Exo4D and QEVD provide actionable feedback commentary for the videos, which consists of timestamped positive and negative critiques, ideas to correct form, pacing, or other aspects of execution. Each sport is critiqued by expert coaches or players from that same sport. Given a video V, the expert pauses the video at multiple timepoints t and provides verbal feedback, e.g., *"...here he's showing good control but lacks speed, which is critical to an effective dribble..."*. Ego-Exo4D additionally has proficiency labels corresponding to four distinct skill levels, from novice to late expert. See Figure 5 and Supp. for examples.

3.3 Stage I: Discovering skill-attributes for pretraining supervision

To supervise the training of $\mathcal{F}_a$, we obtain the skill-attributes set S for every video demonstration by sourcing it from Ego-Exo4D and QEVD commentaries—totaling around $34k$ unique feedback strings. We use these commentaries as a signal to extract the skill-attributes the learner in the video should improve on, as detailed next. While the Ego-Exo4D and QEVD expert commentary represent existing video-language datasets nicely suited for our purposes, such commentary could potentially be sourced "in the wild" from current video sharing platforms (e.g., Reddit, TikTok, etc.) where people provide similar verbal assessments for videos shared on social media. See Supp.

For each training sample, we prompt a large language model (LLM) [2] to extract skill-attributes that are suboptimally demonstrated according to its expert commentary text T. We provide examples to help define the intent of skill-attributes to the LLM. See Supp. for full prompt. This process yields $S = \{s_1, s_2, ...\}$ from the expert commentary T at time t. Note that $|S|$ can be of any length. Fig. 2 shows word clouds of the discovered skill-attributes for the two video datasets. We observe that many salient phrases transcend sport boundaries, like *body positioning*, *balance*, *control*, and *movement*.

Finally, we sample a video chunk around the time t, i.e., $[t - \mu_1, t + \mu_2]$ and associate it with the skill-attribute set S.[1] By using an LLM combined with the raw commentary data, we obtain diverse, open-vocabulary skill-attributes for every training sample. We draw on these resulting LLM-inferred annotations when training the multimodal LLM (Sec. 3.4).

3.4 Stage II: Skill assessment from video

Next we present our model design for learning $\mathcal{F}_a$. Taking inspiration from the success of multimodal LLMs [44, 46, 53, 56], we convert a query video V into visual tokens. Next, we input the visual

[1]The commentary text has only a timepoint t, not a temporal extent; this window extraction is consistent with how the temporal ambiguity is handled in prior work [9, 51, 76].

https://doi.org/10.52202/085713-5359

tokens to a multimodal LLM, along with a prompt to predict the skill-attributes S. The output is parsed to create the prediction $\hat{S}$. Fig. 3 shows the overview, and we describe each step next.

Encoding video demonstrations. We first use a standard video encoder f_v (we use EgoVLP2 [74] and CLIP [75], though others are possible) to extract video features from the video demonstration, $\mathbf{v}' = f_v(V)$, see Fig. 3 (top left). Our model supports either single-view or multi-view inputs. When available, multi-view observations of a skilled physical activity can give valuable detail; for example, an egocentric view of the basketball being shot by the hands along with one or more exocentric viewpoints of the player's full body pose would provide complementary information. When using multiple views, $\mathbf{v}'$ is simply the concatenation of features from individual views. The video spans $\mu_1 + \mu_2$ seconds, as discussed in Sec. 3.3, and one feature is extracted per second. Therefore, $\mathbf{v}'$ is a vector of dimension $(\mu_1 + \mu_2) \times N$ where N is the output feature dimension.

Note that our representation is video-frame based, meaning the person's body pose is visible in the RGB but not explicitly extracted. While one could alternatively supplement the input with explicit body poses, our early experiments showed that the accuracy advantage is minimal compared to the high compute needed to get state-of-the-art poses [79], i.e., 34 seconds for a 10 second clip.

Pretraining: Multimodal LLM for skill-attribute generation. After encoding the videos, we input it to a multimodal LLM (M-LLM), denoted as $\mathcal{L}$. An LLM is a suitable choice since the output attributes are expressed in text format. We prompt the model as follows: *"<video> Here is a video of a person doing <sport name>. Highlight up to <k> key concept areas where the person can improve: ..."*. We replace the *'<video>'* tag with video encodings obtained above.

The output of the M-LLM contains 'k' skill-attributes, distinct phrases that represent the dimensions of suboptimal execution in the provided video demonstration. These outputs are parsed to obtain the skill-attributes set $\hat{S}$, see Fig. 3 (bottom left). To this end, we employ a trainable mapper f_m that converts the feature vector $\mathbf{v}'$ to a representation compatible with the LLM, $\mathbf{v} = f_m(\mathbf{v}')$. The idea is that f_v is a large pretrained model and kept frozen, while f_m is trainable to convert the visual features to a multimodal representation. The supervising signal to train this M-LLM is the skill-attribute set S obtained in Sec. 3.3. We use standard log-likelihood loss [30, 53, 63, 86]. We emphasize that this pretraining enables the model to *generate* skill-attributes, as opposed to retrieving from a closed set. We show the superiority of this approach over retrieval in the experiments.

This pretraining stage results in aligning the video features $\mathbf{v}$ towards skill-attributes, thus making it suitable for the other axes of assessment—actionable feedback and proficiency estimation. We now use the pretraining weights for completing the assessment task suite, as described next.

Skill-attributes for actionable feedback. For generating expert actionable feedback, we provide the output skill-attributes $\hat{S}$, along with video V, as input to the model $\mathcal{L}$ and output the actionable feedback. See Fig. 3 (bottom middle). Next, we prompt the model $\mathcal{L}$ with a prompt as follows: *"<video> Here is a video of a person doing <sport name>. Here are some possible axes that need improvement, as rated by an AI coach (may contain mistakes): <$\hat{S}$>. Give feedback on the execution that will help the person improve."* As above, <video> is replaced by $\mathbf{v}$. Importantly, this prompt provides both V and $\hat{S}$ as an additional guiding signal for the actionable feedback generation process. The generated actionable feedback T contains ideas for improvement that are personalized to the video specifics, e.g., *"you need to bend more while dribbling to maintain control"* as opposed to simply naming the skill dimension to improve, e.g., *control*. While skill-attributes can be the same for two sports, the resulting actionable feedback will be geared towards the specific sport. This factoring helps make our model transferable, while also enabling sport-specific actionable feedback.

Skill-attributes for proficiency estimation. Lastly, we deploy skill-attributes for proficiency estimation. As we want to discern how the skill-attributes representation compares to standard video features without such training, we employ a linear probe setting where we have a linear layer f_p that takes in frozen $\mathbf{v}$ and outputs a class representing the proficiency P, see Fig. 3 (bottom right). The proficiency is a label of expertise: *novice*, *intermediate*, *expert*, or *late expert*. Only f_p is trainable.

3.5 Training and inference settings and implementation details

Train/test splits. Our approach aims to improve both traditional in-domain and zero-shot skill assessment. The datasets organize video clips by their superclass sport and their subclass skill. A skill is a drill or specific exercise, and each sport can have multiple skills. Ego-Exo4D has 3

Table 1: **Quantitative results.** Skill-attribute generation results (IoU@0.7) for Ego-Exo4D [37] and QEVD [68] (top left). Actionable feedback generation results for Ego-Exo4D [37] (top right) and QEVD [68] (bottom left). Metrics used to match SOTA on the corresponding dataset, B@4=BLEU@4, M=Meteor, R-L=ROUGE-L, B=BERT score. Proficiency estimation for individual sports (bottom right). Standard errors are reported in text.

Skill-attribute generation		
Method IoU@0.7	Ego-Exo4D [37]	QEVD [68]
InternVideo2-NN [88]	14.0	23.8
InternVideo2-FT [88]	15.0	24.5
VideoChat2 [47]	9.3	16.9
LLaVA [53]	9.7	17.3
LLaVA-FT [53]	14.6	26.9
Stream-VLM [68]	14.5	28.0
ExpertAF [9]	15.0	28.1
Attribute-Retrieval	19.7	32.4
CrossTrainer	**25.7**	**37.6**

Actionable feedback on Ego-Exo4D [37]			
Method	B@4	M	R-L
InternVideo2-NN [88]	42.1	46.9	49.3
InternVideo2-FT [88]	42.9	47.6	50.0
VideoChat2 [47]	27.8	44.3	41.9
LLaVA [53]	28.5	44.1	44.2
LLaVA-FT [53]	43.5	48.5	51.5
LLaVA-FT w/ pose [53]	43.6	48.5	51.7
PoseScript/Fix [23, 24]	24.1	44.5	46.3
ExpertAF [9, 63]	44.9	49.6	54.6
CrossTrainer	**45.6**	**51.7**	**57.8**
w/o two-stage	43.8	48.8	52.3

Actionable feedback on QEVD [68]			
Method	M	R-L	B
Socratic-LLaMA-2-7B	9.4	7.1	86.0
Video-ChatGPT [56]	10.8	9.3	86.3
LLaMA-VID [50]	10.6	9.0	86.0
Stream-VLM [68]	12.7	11.2	86.3
CrossTrainer	**17.6**	**18.1**	**87.8**
w/o two-stage	12.1	10.8	86.0

Proficiency estimation on Ego-Exo4D [37]			
Method	B.ball	Soccer	Rock Cl.
EgoVLPv2 [74]	48.0	62.5	34.0
CrossTrainer	**53.1**	**68.8**	**37.1**

superclasses and 5 subclasses: soccer has skills dribbling and penalty kick; basketball has skills Mikan layup, reverse layup, jump shot; rock climbing is not sorted into skills. QEVD has 1 superclass (fitness) and 23 subclass skills (jumping jacks, squats, etc.). To explore models' generalization ability, we perform controlled experiments with the following train/test settings, in decreasing volume of available training data (see Fig. 3 (right)):

- *Fully supervised (FS):* Train on all sports and skills, and test on held-out set of videos. This represents in-domain testing.
- *All sport zero-shot (ZS-1):* Train on all sports and skills, *except* the target skill. This means other skills from the same sport are seen during training.
- *Familiar sport zero-shot (ZS-2):* Train on all skills from the same sport, *except* the target skill. This means only the same sport is seen during training.
- *Novel sport zero-shot (ZS-3):* Train only on $n-1$ sports, test on skills from the n-th unseen sport. This means other skills from the same sport are *not* seen during training.

For each controlled experiment, we retrain the model to avoid any information leak between the train and the test splits, and between the seen and novel sports.

Model architecture. f_v is EgoVLPv2 [74] for Ego-Exo4D [37] (precomputed with dataset) and CLIP [75] for QEVD [68] (lightweight to extract). We use one feature per second with a 4096-d output. For Ego-Exo4D [37], we use the ego and four exo views, while QEVD [68] is single-view. f_m is a two layered MLP with GELU activation, consistent with [52]. The MLP+GELU module takes in 4096-d representation, consistent with the embedding dimension of the multimodal LLM $\mathcal{L}$, Llama-3.1-8B-Instruct [5]. f_p is a linear layer with output size 4, the number of proficiency levels.

Training details. We train both $\mathcal{F}_a$ and $\mathcal{F}_t$ in LoRA setting [38], with rank 128, alpha 256, and dropout 0.05, for efficiency. The best performance is obtained with a learning rate of 2×10^{-3} for f_m and 2×10^{-4} for $\mathcal{L}$. Recall that f_v is kept frozen. The model is trained for 2 epochs or till convergence. Total training time depends on the dataset setting, varying between $1-3$ hours. All experiments are performed on one GH200 NVIDIA node.

https://doi.org/10.52202/085713-5359

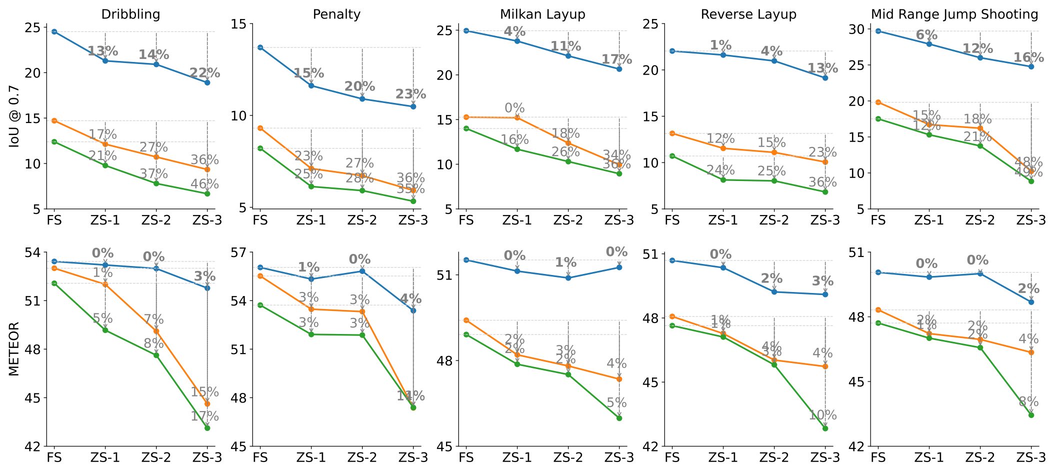

Figure 4: **Zero-shot performance.** Performance trend when testing on various skills (dribbling, penalty, etc.) for different in-domain and zero-shot training settings (FS, ZS-1, etc.) for skill-attribute generation (top) and actionable feedback generation (bottom) for Ego-Exo4D. The relative drop in performance w.r.t. FS is shown as a percentage. Our method is consistently the best for all methods, and the relative drop is the least for all zero-shot variants (ideal curve would be flat and high). See Supp. for QEVD. Legend: —— CrossTrainer, —— ExpertAF [9] and —— InternVideo2 [88].

4 Experiments and Results

We first present actionable feedback generation (Sec. 4.1) for its focus in prior work and directly comparable SOTA methods. Next, we validate the skill-attributes pretraining (Sec. 4.2). Finally, we present proficiency estimation (Sec. 4.3) and explore long-tail in-the-wild YouTube videos (Sec. 4.4).

4.1 Generating actionable feedback

First we evaluate actionable feedback generation: given a video, generate the natural language commentary explaining what to correct.

Baselines. We compare against the SOTA ExpertAF [9] and Stream-VLM [68] models as well as the original baselines from their respective experiments, which include strong multimodal LLMs with model size 7B-8B. CrossTrainer has fewer trainable parameters due to our use of LoRA [38], but we leave the baselines in their full (non-LoRA) form to report their most accurate numbers. "w/o two-stage" is the end-to-end ablation not using skill-attributes.

Metrics. Following [9, 68], we report language metrics—METEOR [11], BLEU-4 [69], ROUGE [78] and BERT-score [25], reported out of 100; higher is better. While establishing ground truth commentary is nuanced and there could be multiple valid feedback statements for a given video, prior work [9] shows that this evaluation paradigm correlates strongly with human subjects' evaluation of the generated commentary.

Fully supervised in-domain results. Tab. 1 (top right and bottom left) shows the results. CrossTrainer clearly outperforms all prior work and strong baselines. Standard error is less than 0.1 for all metrics. This result clearly supports using skill-attributes as an intermediate representation for actionable feedback. We investigate the robustness of actionable feedback generation to the choice of the LLM $\mathcal{L}$ and the noise level in the Supp. Fig. 5 (first two rows) shows qualitative outputs for both datasets, where our model correctly captures the expert's feedback.

Zero-shot transfer. Fig. 4 (bottom row) shows the trend when transferring the knowledge from one domain to another in Ego-Exo4D [37] (see Supp. for QEVD [68]; results are similar). We plot the best M-LLM (ExpertAF [9]) and retrieval (InternVideo2-FT [88]) baselines; we see a similar trend in all other baselines. Our method obtains the best performance for all training settings and degrades most gracefully as training data coverage diminishes in the increasingly difficult zero-shot settings (ZS-1, ZS-2, ZS-3). Our max drop is 4%, vs. 17% for the baseline. In Fig. 5 (bottom left), we show a confusion matrix denoting better transfer from soccer to basketball, and vice versa,

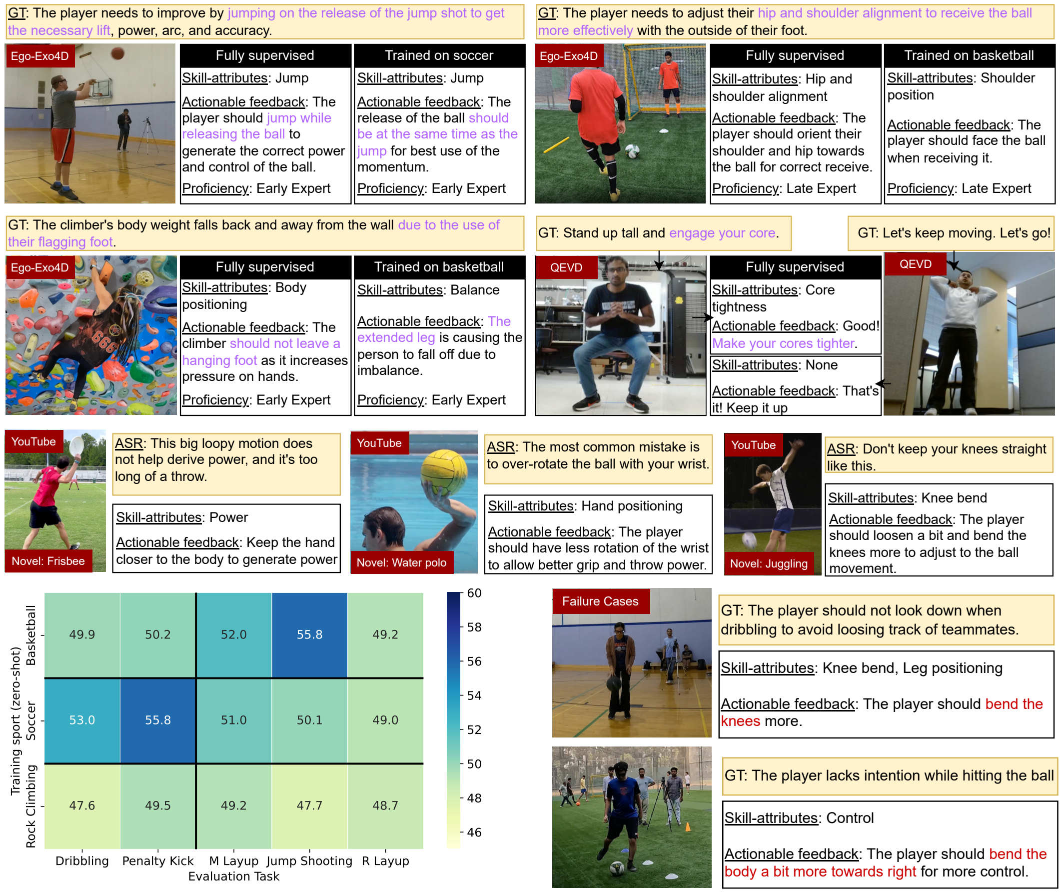

Figure 5: **Qualitative results.** CROSSTRAINER generates skill-attributes, actionable feedback, and proficiency for samples from both Ego-Exo4D [37] and QEVD [68]. The outputs are meaningful even in the zero-shot setting (first two rows). Our method is also applied to in-the-wild videos from YouTube with novel sports (frisbee and water polo) and even new drills (juggling in soccer) with feedback matching the YouTube expert's comments (transcribed with ASR) (third row). Confusion matrix shows better transfer between related sports (bottom left). **Failure cases** here and in Supp. show the difficulty of the task, especially non-visual feedback like *lacking intent* (bottom right).

compared to rock climbing—reinforcing observations in cognitive science [43, 65, 81]. In summary, CROSSTRAINER demonstrates robust to transfer to novel sports and drills.

4.2 Learning skill-attributes in pretraining

Next we evaluate the skill-attribute prediction (c.f. Sec. 3.3).

Baselines. We compare to retrieval baselines (InternVideo2 [88], nearest neighbor (NN) and finetuned (FT), and Attribute-Retrieval trained with contrastive learning), zero-shot multimodal baselines (VideoChat2 [47], LLaVA [53]), and a finetuned version of the best multimodal baseline (LLaVA-FT). In addition, we again compare to ExpertAF [9], a SOTA model for generating actionable feedback, but here we convert its output to skill attributes by text-only LLM prompting [2]. All the baselines have access to the same actionable feedback data as ours [37, 68], and they directly supervise with it.

Metrics. Since skill-attributes are a list of phrases, we employ IoU-based matching (a.k.a. Jaccard index) [21, 31] between the inferred annotations (Sec. 3.3) and generated skill-attributes. We call a pair a match if BERT-score is more than k, and report IoU@k for k=0.7 here and $\{0.8, 1.0\}$ in Supp.

Fully supervised in-domain results. Tab. 1 (top left) shows the results for Ego-Exo4D [37] and QEVD [68]. CROSSTRAINER outperforms all zero-shot and trained baselines that *post process* actionable feedback to obtain skill-attributes, by more than 10%. Moreover, our method outperforms Attribute-Retrieval by 6%—showcasing the effectiveness of our generative approach. The standard

https://doi.org/10.52202/085713-5359

error is less than 0.5 for all metrics and all methods. Fig. 5 shows examples predicting the skill-attributes that the person should improve.

Zero-shot transfer. Fig. 4 (top row of plots) shows the zero-shot transfer results for Ego-Exo4D [37] (see Supp. for QEVD [68]; results are similar). As above, we see a clear advantage of our training scheme. Moreover, CROSSTRAINER declines much more gracefully than the baselines. Overall, these results show the effectiveness of our skill-attribute guided pretraining for zero-shot transfer.

4.3 Estimating demonstrator proficiency

Next we evaluate the quality of the learned video representation $\mathbf{v}$ for proficiency estimation.

Baselines and metrics. We compare the strength of the video representation with respect to the frozen representation f_v, i.e., EgoVLPv2 [74] for all scenarios. We report classification accuracy.

Results. Tab. 1 (bottom right) shows the results. We see a clear gain of up to 6% when using the pretrained video representation, with standard error <1. Fig. 5 shows some example predictions, where our model can distinguish between *early* and *late expert*.[2] This experiment suggests another potential use case in pretraining skill-centric representations for better assessment and feedback.

4.4 Testing long tail in-the-wild YouTube sports videos

Finally, we explore applying CROSSTRAINER to novel sports and drills in in-the-wild videos from YouTube. We extract 10 clips from tutorial videos where an expert tutor demonstrates the incorrect way of doing a drill, while also explaining the incorrectness (i.e., the target feedback, which is withheld from our model). Details of obtaining the videos from YouTube are given in the Supp. We use the expert's transcribed ASR text to compare with the output of our model.

Results. Fig. 5 (fourth row) shows example results (rest in Supp.). We see that a model trained on Ego-Exo4D [37] videos (soccer, basketball, rock climbing), is able to predict the issue with demonstrations in novel long-tail sports frisbee and water polo. Moreover, we also see correct feedback in a novel juggling drill in soccer, which CROSSTRAINER is not trained with. We observe that learning a sport-specific vocabulary is difficult in zero-shot transfer. Nonetheless, the essence is captured and described in text, e.g., even though the model does not understand the *loopy motion*, it understands that keeping the hands closer to the body helps in generating more power—a fact known and applied in various physical scenarios. We also verify the generations with a user study. Human subjects not associated with the project rated 75% of the generations as actionable and correct. Every generation is judged by three raters, and we take a majority vote.

Failure cases in Fig. 5 (bottom right) showcase the difficulty in capturing subtle mistakes, especially non-visual feedback like *intent*, *decision*. We further discuss limitations and societal impact in Supp.

5 Conclusion

We introduced CROSSTRAINER—a novel approach that discovers *skill-attributes* from video demonstrations. These skill-attributes are generalizable, enabling a zero-shot transfer of skill assessment to novel sports and drills. Our experiments show notable gains in actionable feedback generation and proficiency estimation for both in-domain and zero-shot settings. In the future, we will explore ways to quantify sport relatedness to predict the transferability between sports, as well as explicit representations to capture the environmental context.

Acknowledgements

Thanks to the anonymous NeurIPS reviewers for their valuable feedback. This research is supported in part by the UT Austin IFML AI Institute. Compute is from the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at the University of Texas at Austin.

[2]Note that QEVD does not provide proficiency labels.

References

[1] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what?-anticipating temporal occurrences of activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5343–5352, 2018. 3

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 5, 9, 19

[3] Triantafyllos Afouras, Effrosyni Mavroudi, Tushar Nagarajan, Huiyu Wang, and Lorenzo Torresani. Ht-step: Aligning instructional articles with how-to videos. In *NeurIPS*, 2023. 3

[4] Mistral AI. Mistral 8b: A large language model. Online model release, 2024. Version vX.Y (check exact version number) under license (e.g., Apache-2.0). Available at: https://models.mistral.ai/mistral-8b. 20

[5] AI@Meta. Llama 3 model card, 2024. 7, 19

[6] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078, 2023. 3

[7] Kumar Ashutosh, Santhosh Kumar Ramakrishnan, Triantafyllos Afouras, and Kristen Grauman. Video-mined task graphs for keystep recognition in instructional videos. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[8] Kumar Ashutosh, Zihui Xue, Tushar Nagarajan, and Kristen Grauman. Detours for navigating instructional videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18804–18815, 2024. 3

[9] Kumar Ashutosh, Tushar Nagarajan, Georgios Pavlakos, Kris Kitani, and Kristen Grauman. ExpertAF: Expert actionable feedback from video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. 1, 2, 3, 5, 7, 8, 9, 20, 21

[10] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 3

[11] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 8

[12] Gedas Bertasius, Hyun Soo Park, Stella X Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proceedings of the IEEE international conference on computer vision*, pages 2177–2185, 2017. 1, 2, 3, 21

[13] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 3

[14] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 3

[15] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[16] James Burgess, Xiaohan Wang, Yuhui Zhang, Anita Rau, Alejandro Lozano, Lisa Dunlap, Trevor Darrell, and Serena Yeung-Levy. Video action differencing. *arXiv preprint arXiv:2503.07860*, 2025. 1

[17] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[18] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 1

[19] Joe Causer and Paul R Ford. "decisions, decisions, decisions": transfer and specificity of decision-making skill between sports. *Cognitive Processing*, 15:385–389, 2014. 2, 3

https://doi.org/10.52202/085713-5359

[20] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI*, pages 334–350. Springer, 2020. 3

[21] Yunmo Chen, William Gantt, Tongfei Chen, Aaron Steven White, and Benjamin Van Durme. A unified view of evaluation metrics for structured prediction. *arXiv preprint arXiv:2310.13793*, 2023. 9

[22] CVsports, 2025. 3

[23] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, pages 346–362. Springer, 2022. 7

[24] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: Correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023. 7

[25] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 8, 20

[26] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 3

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[28] Hazel Doughty, Dima Damen, and Walterio Mayol-Cuevas. Who's better? who's best? pairwise deep ranking for skill determination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[29] Hazel Doughty, Walterio Mayol-Cuevas, and Dima Damen. The pros and cons: Rank-aware temporal attention for skill determination in long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[30] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6

[31] Sam Fletcher, Md Zahidul Islam, et al. Comparing sets of patterns with the jaccard index. *Australasian Journal of Information Systems*, 22, 2018. 9

[32] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020. 3

[33] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 3

[34] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1711–1721, 2018. 3

[35] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 3

[36] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16102–16112, 2022. 3

[37] 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. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 2, 3, 4, 7, 8, 9, 10, 18, 20, 21, 27

[38] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 7, 8

[39] Yifei Huang, Guo Chen, Jilan Xu, Mingfang Zhang, Lijin Yang, Baoqi Pei, Hongjie Zhang, Lu Dong, Yali Wang, Limin Wang, et al. Egoexolearn: A dataset for bridging asynchronous ego-and exo-centric view of procedural activities in real world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22072–22086, 2024. 1

[40] Dat Huynh, Jason Kuen, Zhe Lin, Jiuxiang Gu, and Ehsan Elhamifar. Open-vocabulary instance segmentation via robust cross-modal pseudo-labeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7031, 2022. 3

[41] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NeurIPS*, 2014. 3

[42] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, 2008. 3

[43] Daniel Leach, Zoe Kolokotroni, and Andrew D Wilson. Perceptual information supports transfer of learning in coordinated rhythmic movement. *Psychological Research*, 85(3):1167–1182, 2021. 3, 4, 9

[44] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*, 2023. 5

[45] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *arXiv preprint arXiv:2201.09450*, 2022. 3

[46] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 5, 28

[47] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark, 2023. 7, 9, 20

[48] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 3

[49] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 3

[50] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 7

[51] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3, 5

[52] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 7

[53] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 5, 6, 7, 9, 20, 28

[54] Jingen Liu, Ben Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 3

[55] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 1

[56] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 5, 7

[57] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action forecasting@ ego4d challenge 2022. *arXiv preprint arXiv:2207.12080*, 2022. 3

[58] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations, 2023. 3

 https://doi.org/10.52202/085713-5359

[59] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 3

[60] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9879–9889, 2020. 1

[61] Tushar Nagarajan and Lorenzo Torresani. Step differences in instructional video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[62] Zwe Naing and Ehsan Elhamifar. Procedure completion by learning from partial summaries. In *British Machine Vision Conference*, 2020. 3

[63] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 6, 7

[64] OpenFAD, 2023. 3

[65] Luca Oppici and Derek Panchuk. Specific and general transfer of perceptual-motor skills and learning between sports: A systematic review. *Psychology of Sport and Exercise*, 59:102118, 2022. 2, 9

[66] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *NeurIPS*, 2009. 3

[67] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. *arXiv preprint arXiv:2503.20781*, 2025. 1, 3

[68] Sunny Panchal, Apratim Bhattacharyya, Guillaume Berger, Antoine Mercier, Cornelius Böhm, Florian Dietrichkeit, Reza Pourreza, Xuanlin Li, Pulkit Madan, Mingu Lee, et al. What to say and when to say it: Live fitness coaching as a testbed for situated interaction. *Advances in Neural Information Processing Systems*, 37:75853–75882, 2024. 1, 3, 4, 7, 8, 9, 10, 18, 19, 20, 21, 27

[69] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 8

[70] Paritosh Parmar and Brendan Morris. Action quality assessment across multiple actions. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1468–1476, 2019. 1, 2, 3

[71] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 3

[72] Paritosh Parmar, Amol Gharat, and Helge Rhodin. Domain knowledge-informed self-supervised representations for workout form assessment. In *European Conference on Computer Vision*, pages 105–123. Springer, 2022. 1, 2

[73] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 556–571. Springer, 2014. 1, 3

[74] Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Zheng Shou, Rama Chellappa, and Pengchuan Zhang. Egovlpv2: Egocentric video-language pre-training with fusion in the backbone. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5285–5297, 2023. 3, 6, 7, 10

[75] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 6, 7

[76] Santhosh Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Naq: Leveraging narrations as queries to supervise episodic memory. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[77] André Roca and A. Mark Williams. Does decision making transfer across similar and dissimilar sports? *Psychology of Sport and Exercise*, 31:40–43, 2017. 2

[78] Lin CY Rouge. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 8

[79] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024. 6

[80] Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. Howtocaption: Prompting llms to transform video annotations at scale. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 3

[81] Winona Snapp-Childs, Andrew D Wilson, and Geoffrey P Bingham. Transfer of learning between unimanual and bimanual rhythmic movement coordination: Transfer is a function of the task dynamic. *Experimental brain research*, 233(7):2225–2238, 2015. 3, 4, 9

[82] Ben William Strafford, Pawel Van Der Steen, Keith Davids, and Joseph Antony Stone. Parkour as a donor sport for athletic development in youth team sports: Insights through an ecological dynamics lens. *Sports medicine-open*, 4:1–6, 2018. 2, 3, 4

[83] Tristan Sylvain, Linda Petrini, and Devon Hjelm. Locality and compositionality in zero-shot learning. In *ICLR*, 2020. 3

[84] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 3

 https://doi.org/10.52202/085713-5359

[85] Yansong Tang, Zanlin Ni, Jiahuan Zhou, Danyang Zhang, Jiwen Lu, Ying Wu, and Jie Zhou. Uncertainty-aware score distribution learning for action quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9839–9848, 2020. 1, 2, 3

[86] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. 6

[87] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022. 1, 3

[88] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 1, 3, 7, 8, 9, 19, 20

[89] Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13587–13597, 2022. 3

[90] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018. 3

[91] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 3

[92] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 1, 3

[93] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 3

[94] Wenjia Xu, Yongqin Xian, Jiuniu Wang, Bernt Schiele, and Zeynep Akata. Attribute prototype network for zero-shot learning. In *NeurIPS*, 2020. 3

[95] Xun Xu, Timothy Hospedales, and Shaogang Gong. Semantic embedding space for zero-shot action recognition. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 63–67. IEEE, 2015. 3

[96] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. *arXiv preprint arXiv:2212.04979*, 2022. 3

[97] Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, and Cordelia Schmid. Unloc: A unified framework for video localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13623–13633, 2023. 3

[98] Xumin Yu, Yongming Rao, Wenliang Zhao, Jiwen Lu, and Jie Zhou. Group-aware contrastive regression for action quality assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7919–7928, 2021. 1, 2, 3

[99] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020. 3

[100] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12870–12877, 2020. 3

[101] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. *arXiv preprint arXiv:2303.17839*, 2023. 3

[102] Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10727–10738, 2023. 3

[103] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 3

https://doi.org/10.52202/085713-5359

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 [37], QEVD [68], 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.4: Additional quantitative** results to show metrics for skill-attribute generation, and QEVD [68] zero-shot results.
- We discuss **limitations** in **Sec. A.5**.
- We also discuss **societal impact** in **Sec. A.6**.

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 commentating 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 Obtaining YouTube videos for testing long tail in-the-wild sports and drills

We evaluate the performance of our model on zero-shot sports and drills from in-the-wild YouTube videos in Sec. 4.4. To obtain a dataset for this test, we create a list of novel sports and novel drills within the Ego-Exo4D [37] and QEVD [68] sports (basketball, soccer, rock climbing, exercise), and search for videos on YouTube with their coaching videos. Our criteria is that the video should contain a mistake done by a learner, and a coach giving feedback. Note that many videos show only the correct demonstration; hence, obtaining videos for our task is challenging.

Specifically, we randomly selected some common sports—soccer (juggling), basketball (dribbling) and some rarer ones like water polo, korfball, polo, jai alai, kin ball, frisbee. We search more than 50 videos with keywords like "Coaching video of <sport/drill>", "<sport/drill> training session for beginners", "<sport/drill> common mistakes for beginners", "Dos and don'ts in <sport/drill>". The videos are manually watched to find the desired coaching instances. The process took overall 12 hours, and was done by graduate students not associated with this project.

https://doi.org/10.52202/085713-5359

Table 2: Buckets of exercises in QEVD [68] 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.

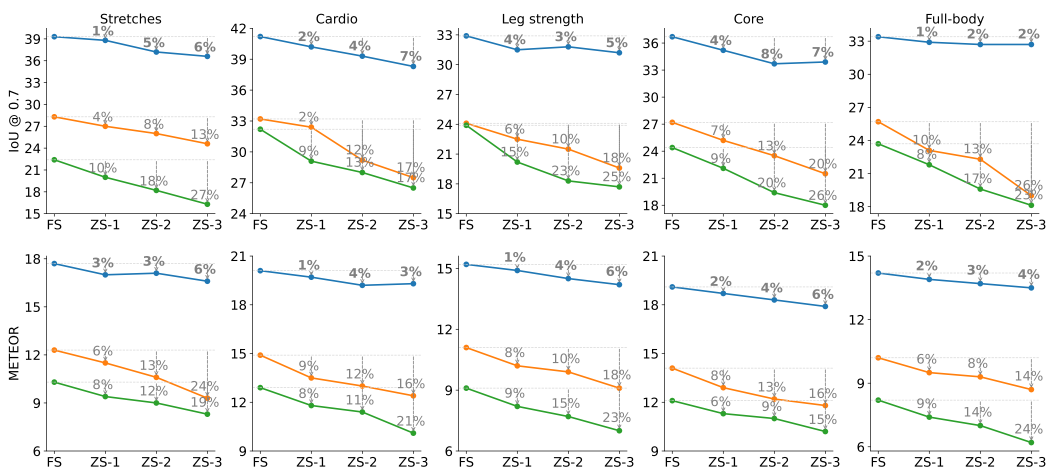

Figure 6: **Zero-shot performance.** Performance trend when testing on various skills (stretches, cardio, etc.) for different in-domain and zero-shot training settings (FS, ZS-1, etc.) for skill-attribute generation (top) and actionable feedback generation (bottom) for QEVD [68]. Legend: —— CrossTrainer, —— Stream-VLM [68] and —— InternVideo2 [88].

A.4 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 [68] 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 [5], 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.

https://doi.org/10.52202/085713-5359

Table 3: **Quantitative results.** Skill-attribute generation results for IoU@k for $k \in \{0.7, 0.8, 1.0\}$ for Ego-Exo4D [37] and QEVD [68], extension of Tab. 1 (top left) on remaining k values.

Method	IoU @0.7	@0.8	@1.0
InternVideo2-NN [88]	14.0	8.9	7.7
InternVideo2-FT [88]	15.0	9.4	8.2
VideoChat2 [47]	9.3	6.6	4.0
LLaVA [53]	9.7	7.2	4.9
LLaVA-FT [53]	14.6	9.1	8.1
Stream-VLM [68]	14.5	9.1	8.3
ExpertAF [9]	15.0	9.5	8.4
Attribute-Retrieval	19.7	12.7	10.7
CROSSTRAINER	**25.7**	**15.9**	**14.4**

Method	IoU @0.7	@0.8	@1.0
InternVideo2-NN [88]	23.8	16.3	14.8
InternVideo2-FT [88]	24.5	16.6	15.3
VideoChat2 [47]	16.9	11.4	10.1
LLaVA [53]	17.3	12.5	11.8
LLaVA-FT [53]	26.9	19.2	18.2
Stream-VLM [68]	28.0	19.9	18.6
ExpertAF [9]	28.1	19.7	18.3
Attribute-Retrieval	32.4	24.7	23.0
CROSSTRAINER	**37.6**	**29.8**	**28.1**

Table 4: **Robustness and sensitivity of LLM:** Comparison of actionable feedback generation with various sources of skill-attributes (left). Comparison of BERT similarity score of skill-attributes generated by our method with skill-attributes obtained from different LLMs and prompts (right).

Skill-attrib. source	**Test-set $\mathcal{L}$**	**B@4**	**M**	**R-L**
Llama-3 8B (orig)	Llama-3 8B (orig)	45.6	51.7	57.8
Mistral 8B [4]	Llama-3 8B (orig)	45.3	51.8	57.5
Llama-3 8B (orig)	Mistral 8B	45.8	51.8	57.8
Llama-3 8B (orig)				
w/ 10% noise	Llama-3 8B (orig)	45.2	50.5	56.2
w/ 20% noise	Llama-3 8B (orig)	45.1	50.0	54.9
w/ 30% noise	Llama-3 8B (orig)	44.3	49.4	53.1
w/ 50% noise	Llama-3 8B (orig)	42.7	47.3	50.2
w/ 70% noise	Llama-3 8B (orig)	41.9	46.3	49.6

Ours vs	**Score**
Llama-3 8B w/ prompt choice 1	0.99
Llama-3 8B w/ prompt choice 2	0.98
Mistral 8B [4]	0.98

Fig. 6 shows the results. First, in skill-attribute generation (top row), we see that our method outperforms both Stream-VLM [68] and InternVideo2 [88] baselines. Moreover, as seen in Ego-Exo4D [37], the performance decrease in the zero-shot setting is milder than the drop observed in the baselines. Finally, we see a similar trend in actionable feedback generation (bottom row). Overall, zero-shot results in both Ego-Exo4D [37] and QEVD [68] show that our idea of learning to generalize using skill-attributes is effective.

Finally, to show the robustness and sensitivity, we perform the following experiments on Ego-Exo4D [37] actionable feedback generation:

Robustness to the choice of the language model: We train and evaluate skill-attributes using different language models. We first train the model using Mistral's 8B language model (mistralai/Ministral-8B-Instruct-2410) [4] and compare it against the skill-attributes test set obtained using Llama-3.1-8B-Instruct, and vice-versa. Tab. 4 (left) shows the results. We see that the model is robust to the choice of the language model, and using any strong language model helps achieve a good performance.

Actionable feedback generation w/ noisy skill-attributes: We inject noise in the actionable feedback generation evaluation. We replace X% of inferred skill-attributes with a random skill-attribute and observe the performance at various levels of noise X. See results in the table below. We observe that adding noise degrades the performance, with the performance matching that of end-to-end direct training at $X = 20\%$. We can conclude that the performance is positively correlated to the quality of the generated skill-attribute. Improving that will also improve the actionable feedback performance. These ablation studies further showcase the effectiveness of using skill-attributes for actionable feedback.

Correlation between skill-attributes generated using different LLMs: We try two prompt variants, and a different language model (Mistral 8B, mistralai/Ministral-8B-Instruct-2410) [4] to compare the similarity between generated skill-attributes—checking if our idea is independent of the chosen language model. We use Hungarian matching to find the most-similar match between the new skill-attribute set and our original skill-attribute set. Next, we find the average BERT score [25] between the sets. Tab 4 (right) shows the results. We see a very high similarity between skill-attributes generated from different prompts, and a different language model. This result implies that our idea is independent of the choice of a reasonable language model.

A.5 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.

Moreover, in this work, we do not factor aspects like terrain and opponent behavior. This assumption is consistent with the datasets Ego-Exo4D [37] and QEVD [68] that are both single-person, and do not consider external factors. Furthermore, prior work also considers single-person skill assessment/feedback [9, 12, 37]. There are research works in multi-agent cooperation, but they are restricted to simulation and simpler objectives than performance feedback.

A.6 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 [37] and QEVD [68] 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.

https://doi.org/10.52202/085713-5359

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: [Yes] We show results on transferring performance across various sports on multiple datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

https://doi.org/10.52202/085713-5359

Justification: [Yes] Sec. 4.4 discusses the limitations of our work.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: [NA] We do not propose any theoretical result.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: [Yes] Sec. 3.5 describes all the information required to reproduce the code. Furthermore, we will release the code and the data upon paper acceptance.

   Guidelines:

https://doi.org/10.52202/085713-5359

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: [No] The code and data will be released upon acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: [Yes] The method section (Sec. 3) discusses the training and test details, along with all the design choices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: [Yes] All result sections have the maximum standard error reported.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: [Yes] The details of compute resource and training duration is given in Sec. 3.5. We did not have significant computational overhead due to failed experiments.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

https://doi.org/10.52202/085713-5359

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: [Yes] The authors confirm that this research follows the NeurIPS code of ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: [Yes] Supplementary discusses societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

Justification: [Yes] We only use Exo-Exo4D [37] and QEVD [68] dataset for the research. Moreover, the generations of the language model is finetuned to only generate expert commentary. We believe our trained model cannot be misused.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: [Yes] We have cited and ensured all the codes and models have licenses that we can use for this research.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: [Yes] All the details of the weakly-supervised dataset is provided in Sec. 3. We will provide the documentation of the code along with its release to the community.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

https://doi.org/10.52202/085713-5359

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: [NA]

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: [Yes] Our model design use LLM as the core component. This is a standard in recent video understanding methods [46, 53].

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

https://doi.org/10.52202/085713-5359