# ExpertAF: Expert Actionable Feedback from Video

Kumar Ashutosh[1,2], Tushar Nagarajan[2], Georgios Pavlakos[1], Kris Kitani[2,3], Kristen Grauman[1,2]
[1] University of Texas at Austin, [2] FAIR Meta, [3] Carnegie Mellon University

## Abstract

*Feedback is essential for learning a new skill or improving one's current skill-level. However, current methods for skill-assessment from video only provide scores or compare demonstrations, leaving the burden of knowing what to do differently on the user. We introduce a novel method to generate* actionable feedback *(AF) from video of a person doing a physical activity, such as basketball or soccer. Our method takes a video demonstration and its accompanying 3D body pose and generates (1) free-form expert commentary describing what the person is doing well and what they could improve, and (2) a visual expert demonstration that incorporates the required corrections. We show how to leverage Ego-Exo4D's [29] videos of skilled activity and expert commentary together with a strong language model to create a weakly-supervised training dataset for this task, and we devise a multimodal video-language model to infer coaching feedback. Our method is able to reason across multi-modal input combinations to output full-spectrum, actionable coaching—expert commentary, expert video retrieval, and expert pose generation—outperforming strong vision-language models on both established metrics and human preference studies.*

## 1. Introduction

An abundance of instructional "how-to" videos on the internet enables skill-learning by observing expert demonstrations. Instructional videos cover many skills one might want to learn—cooking, DIY, sports, crafts, and hobbies. Recent work leverages such videos to assist human [5, 6, 53] and robot [56] skill learning. Acquiring skills from video is especially appealing for equitable learning: access to video is generally less costly and more widely available than face-to-face access to an expert coach.

Despite the myriad of videos and tutorials, people prefer to learn in a feedback loop: getting to know their mistakes, finding the improvements they need to do and correcting those. For that reason, coaching-based iterative training is more effective than self-learning [48]. A good coach has three components: finding mistakes, providing verbal corrections and finally, showing visual demonstrations. For example, suppose a person learning to dribble a soccer ball takes big steps that decrease body control. The coach should identify this mistake, provide verbal feedback, and show the correct dribbling technique. See Figure 1.

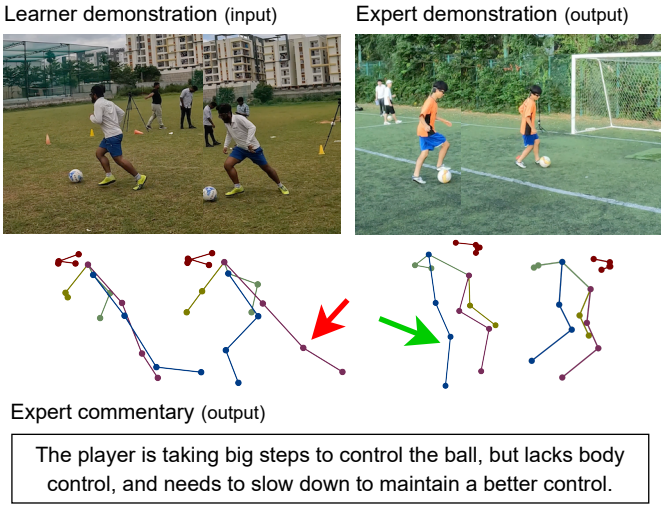


Figure 1. **An example of expert feedback.** When a player is dribbling the ball fast, they tend to lose control (top left). Our proposed method provides an expert commentary to the learner suggesting improvements (bottom). The method also provides an expert demonstration that shows the desired correction, where the player is maintaining smaller steps and body control (top right).

At present, a person watching a how-to video has to identify mistakes themselves and attempt to correct them. Pinpointing the exact mistake is itself challenging for beginners—particularly when learning *physical* skills, like sports and dance, where subtle movements and positioning of the body are critical to success.

What role could AI play in overcoming this gap? Current work on proficiency understanding from video focuses on *scoring* a demonstration [9, 28, 29, 78], thus addressing only one component of coaching. There is no method that can provide actionable verbal and visual feedback on a learner's demonstration. Doing so is technically challenging since it requires the method to (a) understand the user's activity, (b) detect mistakes in the execution by comparing

Project page: https://vision.cs.utexas.edu/projects/ExpertAF/

the user's activity with *correct* examples and finally, (c) suggest edits with respect to the better way of doing the activity. In short, the output should be *coach-like*—specific actionable feedback that will improve a learner's proficiency.

We propose ExpertAF: a novel method to provide expert actionable feedback on a user's physical activity. ExpertAF's actionable feedback has two forms: language-based commentary and visual demonstration. The free-form expert commentary output describes the mistake and what should be corrected, e.g., *"take smaller steps and slow down to maintain control"*, while the expert visual demonstration output shows the correct way of doing it.

Both components introduce unique technical challenges. For the former, the task of generating free-form text feedback is distinct from standard image and video captioning, as the model needs to identify the mistake and actions that would correct it—rather than simply describe the observed pose [16] or activity [32, 38, 57, 75, 81]. Similarly, for the latter, the task of retrieving (or generating) a demonstration that *corrects* a specific mistake is distinct from similarity-based retrieval [3, 40, 54, 77, 79] or open-ended generation [12, 19]. We hypothesize—and experimentally validate—that personalized, video-conditioned coaching is more accurate than simply returning a global expert execution because there are many right ways of doing an activity, e.g., *penalty shots* in soccer, and hence different ways to fix what is wrong in a given execution. Overall, the model must understand both what is being done *and* what actionable tweaks would make it better.

We are the first to address video-based coaching with actionable feedback. We develop the first model to generate free-form text feedback on a user's activity (captured in video and 3D pose sequences) as well as a video demonstrating how to improve, by building on a recent vision-language model [45]. There is currently no dataset for this challenging task. Therefore, we augment Ego-Exo4D [29], an existing dataset with video, commentary, and poses, and show how to use a strong large language model to create a weakly-supervised training set consisting of paired learner-expert demonstrations, along with expert textual commentary that relates the two. We also obtain a gold-standard manually labeled test set for rigorous evaluation.

We validate ExpertAF on three diverse physical scenarios: soccer, basketball, and rock climbing. Our results show a significant improvement over strong baselines, establishing the first method to provide actionable feedback, including a novel technique for expert pose generation. Alongside consistent gains in quantitative metrics against the ground truth, we also show our model's promise via direct human evaluation, where ExpertAF outperforms off-the-shelf video models by as much as 3×.

## 2. Related work

**Skill learning from videos.** Instructional video datasets like HowTo100M [53], COIN [70], and CrossTask [87] enable procedural understanding through step recognition [5, 42, 85], procedure planning [11, 13, 84], task-graph discovery [5, 18, 29, 85], action anticipation [1, 21, 22, 26, 51], and alignment detection [4, 30]. Capitalizing on their instructional nature, recent work learns robot policies from these videos [50, 69], uses external knowledge-bases like WikiHow to ground the steps of a procedural task [42, 52, 85], and explores new ways to navigate between multiple demonstrations [6]. Despite their scale, how-to videos in these datasets [53, 70, 87] are often created by domain experts and hence lack mistakes or suboptimal executions. In contrast, Ego-Exo4D [29] contains multiple executions of various common scenarios by people with varied skill levels—from novice to late expert. Moreover, it contains experts' feedback on those actions. We show how to augment this data via large language models to support our new idea for video-based coaching.

**3D body pose for activity understanding.** 3D body pose is crucial for human activity understanding, capturing a person's stance and motion. Supported by valuable datasets like Humans3.6M [33] and NTU RGB-D [46, 66], recent methods improve body pose understanding [25, 34, 67, 72, 83, 86]. Beyond action classification, pose is also crucial for markerless motion capture of sports [8, 15] and interacting in augmented reality [24, 31]. While our ExpertAF leverages pose as an important signal of human activity, unlike prior work we interface video, pose, and free-form text to generate expert feedback and an expert demonstration video. Our work is also different from generating [20] or modifying [17] body pose from text (e.g.,*"raise your left arm"*), as our task is to understand a potentially incorrect pose and provide feedback. Understanding these minor differences between incorrect and correct poses offers unique challenges that are addressed by our proposed learning scheme. Our work is orthogonal to methods that improve pose estimation itself; any future improvements on that front would only benefit our approach.

**Skill assessment and coaching.** Prior work explores skill assessment for a variety of tasks, particularly sports. Most methods pose skill-assessment as a score-prediction task, i.e., for skating [78], gymnastics [28], basketball [9], or multiple sports [61]. Since absolute scoring can be ambiguous and error-prone, recent work [71] explores uncertainty-aware score distributions. Instead of explicitly scoring a demonstration, other work determines which of two basketball players are better [9], uses group-aware contrastive regression to learn the relative quality of demonstrations [80], or generates a full ideal trajectory from the first frame [10]. Fitness-AQA [63] and Action Quality [64] provide outputs like *knees inward error*, *shallow squat error*,

or an arrow highlight for more localized feedback. There is also research on skill assessment in non-sports domains, like surgery [43, 76] and piano [62]. All the prior work assumes a fixed taxonomy of errors, and the taxonomy is designed separately for each activity. Furthermore, unlike our work, none of the prior work provides feedback akin to a personal coach—which requires not only telling what is wrong, but also expressing how to correct the mistake.

## 3. Method

We introduce the problem statement in Sec. 3.1, the dataset creation strategy in Sec. 3.2, the training design in Sec. 3.3, and implementation details in Sec. 3.4.

### 3.1. Problem statement

Consider a dataset $\mathcal{E} = \{(\mathcal{V}, T, \bar{\mathcal{V}})\}$ where each $\mathcal{V}_i$ is a video demonstration, and $T_i$ is a free-form text commentary critiquing the activity in the video, e.g., *"take smaller steps and slow down to maintain control."* $\bar{\mathcal{V}}$ is a related video demonstration but without the error mentioned in $T$, e.g., the player takes smaller steps with better control. See Fig. 1. Each demonstration $\mathcal{V} = \{V, P, S\}$ consists of three parts: $V$ is the RGB video, and $P$ is the 3D pose sequence of the participant with skill-level $S$ in the video. The pose representation $P \in \mathbb{R}^{n \times d \times 3}$ contains $n$ frames of 3D positions for $d = 17$ body joints, consistent with MS-COCO [41]. $S$ is the participant's skill-level for an activity, i.e., *novice*, *early expert*, *intermediate expert*, or *late expert*, and will be used later in creating the dataset. We assume there is one active person doing a physical activity in any given video who is the subject of ExpertAF.

The goal in this work is to provide feedback on a given learner's video containing a physical scenario. The feedback is an expert demonstration $\bar{\mathcal{V}}$ and an expert commentary $T$, tailored to $\mathcal{V}$. Formally, we want to learn a mapping $\mathcal{V} \rightarrow (\bar{\mathcal{V}}, T)$. While it is possible to learn these mappings separately (i.e., $\mathcal{V} \rightarrow \bar{\mathcal{V}}$ and $\mathcal{V} \rightarrow T$), the expert commentary and demonstration are tightly related and provide important contextual information for generating each other. We therefore treat this as an autoregressive generation problem where during training, we use both $\mathcal{V}$ and $T$ to generate $\bar{\mathcal{V}}$ (or $\mathcal{V}$ and $\bar{\mathcal{V}}$ to generate $T$), while during testing, we drop the extra context information to generate $\bar{\mathcal{V}}$ and $T$ directly.

Mathematically, consider $\mathcal{F} : (\mathcal{V}, \bar{\mathcal{V}}, T) \rightarrow \mathbb{R}$ as the autoregressive function that jointly learns from the learner demonstration $\mathcal{V}$, the expert demonstration $\bar{\mathcal{V}}$, and the expert commentary $T$. The output $\in \mathbb{R}$ is the loss that we aim to minimize. We use the same unified $\mathcal{F}$ for the joint training (detailed in Sec. 3.3) and inference, as follows. We use $\mathcal{F}'$ when the autoregressive model $\mathcal{F}$ is used for output token generation, i.e., text or pose tokens. $\mathcal{F}'$ can take in any input used during training and generate the remaining one—typical for autoregressive models [74].

**Expert commentary generation.** At inference time, to generate an expert commentary $T$, we only have a learner demonstration, thus we mask out the expert demonstration and output the expert commentary:

$$\hat{T} = \mathcal{F}_t(\mathcal{V}) = \mathcal{F}'(\mathcal{V}, \varnothing) \tag{1}$$

where $\mathcal{F}_t$ denotes using the model to generate commentary.

**Expert demonstration generation.** Next, we obtain the expert demonstration in two forms—retrieving a correct execution (video and pose) and generating a pose sequence. These two output formats offer complementary information to understand the actionable feedback. While a video exemplar is helpful for many learner mistakes, pose generation is useful in the absence of a correct expert demonstration in the retrieval set—allowing our model to generalize its coaching beyond the set of discrete expert videos. We leave expert video generation as a future work.

Denoting $\mathcal{F}_r$ and $\mathcal{F}_g$ as the expert demonstration retrieval and expert pose generation functions, respectively, we have:

$$\bar{\mathcal{V}} = \mathcal{F}_r(\mathcal{V}) = \underset{\forall \mathcal{V}' \in \mathcal{E}}{\operatorname{argmin}} \mathcal{F}(\mathcal{V}, \mathcal{V}', \hat{T}) \tag{2}$$

$$P = \mathcal{F}_g(\mathcal{V}) = \mathcal{F}'(\mathcal{V}, \hat{T}) \tag{3}$$

where $\hat{T}$ is the output from Eq. 1 above, and $\mathcal{F}_r$ and $\mathcal{F}_g$ denote using the model to retrieve or generate the expert demonstration, respectively.

In summary, the training step $\mathcal{F}$ uses tuples $\mathcal{V}, \bar{\mathcal{V}}, T$ in a unified way, described in Sec. 3.3, whereas, at inference, all the three functions $\mathcal{F}_t$, $\mathcal{F}_r$ and $\mathcal{F}_g$ use only $\mathcal{V}$.

### 3.2. Forming the expert feedback dataset

To learn the desired functions, we need to obtain pairs of videos $(\mathcal{V}, \bar{\mathcal{V}})$ where there is an error in the demonstration in the first video that is corrected in the second video, along with the expert commentary $T$ about $\mathcal{V}$. Ego-Exo4D [29] offers a great starting point for our setup. It contains ego-exo videos, extracted 3D pose sequences, and time-stamped commentary by experts (e.g., professional soccer coaches) on the demonstrations in the video. The experts watched the entire video and stopped each time they saw something to critique or compliment, offering free-form spoken commentary. In total, this led to 117,812 sentences across 221 hours of video, with most videos commentated by 2-5 unique experts [29]. See examples in Figure 2 (left) and details in Sec. 3.4 and Supp.

We propose to automatically augment this data in two ways to enable coaching. First, we seek (pseudo-)annotations of whether a given commentary statement describes a needed improvement or applauds a correct execution. Second, we localize each piece of feedback on a body region, e.g., *incorrect hand stretch* vs. *wrong legs movement*. These distinctions are crucial to generate feedback

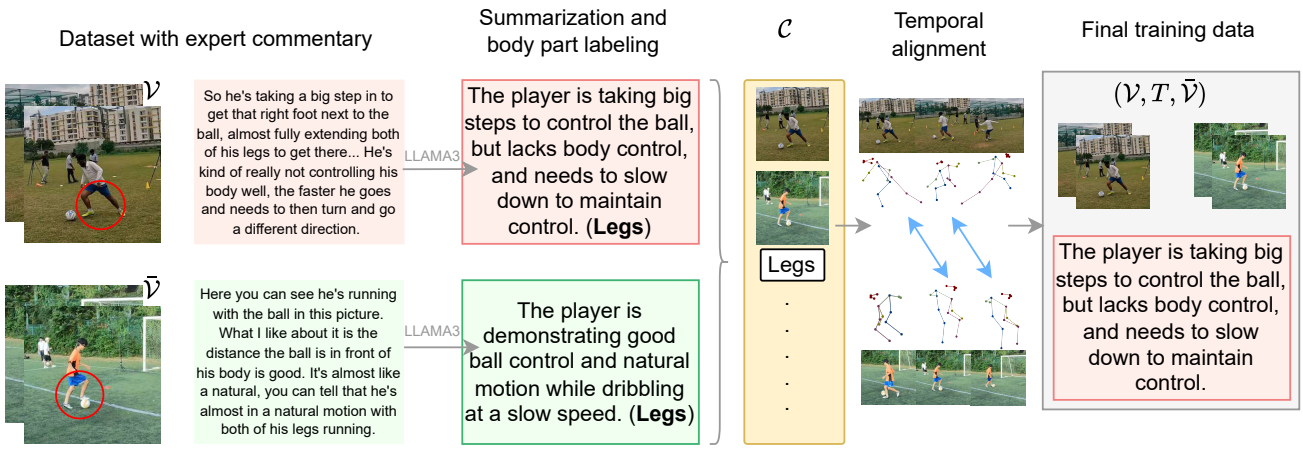


Figure 2. **Overview of the dataset creation.** We first summarize the human-provided expert commentary [29] in one sentence using an LLM, and then map it to a body region and correct (green) or incorrect (red) execution label. We then choose incorrect-correct pairs for the same body region to obtain $\mathcal{C}$. Finally, we choose pairs with minimum temporal alignment loss to obtain the training data. Best in zoom.

and show corrections. To this end, we create a weakly-supervised training dataset consisting of tuples $(\mathcal{V}, T, \bar{\mathcal{V}})$ from Ego-Exo4D [29], as follows.

**Expert commentary classification and body localization.** The commentary in [29] is obtained from voice recordings of the experts converted to text with ASR. Most of the samples contain extra comments like *"oh, I will give this a five out of ten"* and *"that's how I would do it too"*. Additionally, as discussed above, they lack positive and negative labels and do not explicitly indicate the body region involved (e.g., *legs*, *arms*). Thus, we preprocess the expert commentary for three things—making the commentary concise to extract the improvable feedback, marking which body region the feedback is about, and marking whether the feedback states the need for improvement or not.

Since these are all addressable with text reasoning, we use a large language model (LLM) to provide the desired answers. We use Llama3-70B [2], a recent open-sourced model that performs well in current benchmarks. In essence, given a commentary $T$, the language model $\mathcal{L}$ responds to the prompt: *"Given an expert's commentary, summarize the feedback into a single sentence and also provide which body regions need improvement or are correctly executed..."* (see Supp.). Formally, this yields $\mathcal{L}(T) = (T', (b^1, c^j), ..., (b^s, c^s))$ where $T'$ is the concise summary and $b^i$ is a body region like *head* or *arms* and $c^i \in \{0, 1, 2\}$ are labels representing *needs improvement*, *correct execution*, and *no mention*, respectively. We group the skeleton joints into six pre-defined body regions $b^i$ (details in Supp.). See red and green boxes in Fig. 2.

**Pairing incorrect and correct executions.** Next, we use the above information to mine pairs of incorrect and correct execution in the dataset. Ego-Exo4D also contains metadata about the skill-level $S$ of the demonstrator, broken into four categories (1-4 in increasing expertise) starting from *novices* who have not performed the activity before to *late experts* who have performed the same activity, e.g., *basketball* for 10+ years. We sample incorrect learner demonstrations from beginners and correct demonstrations from experts from the same activity in Ego-Exo4D, e.g., penalty kicks in soccer or reverse layup in basketball. Even though there can be incorrect executions by experts and correct ones by beginners, errors by experts are likely incomparable to beginners' due to the skill gap. We use the mapping of body regions to find (in)correct executions referring to the same body region. This results in a collection of video pairs with negative and positive feedback about the same body region, e.g., legs in Fig. 2. Formally, the collection $\mathcal{C}$ is curated as

$$\{(\mathcal{V}_1, T, \mathcal{V}_2) \mid S_1 \in \mathcal{S}^n, S_2 \in \mathcal{S}^e, \exists j \; s.t. \; c_1^j = 0, c_2^j = 1\}$$

where $\mathcal{S}^n$ ={novice, early expert} and $\mathcal{S}^e$ = {late expert, intermediate expert}. Using this matching, we obtain a collection of video pairs of incorrect and correct execution (Fig. 2, yellow box).

**Temporal alignment and filtering demonstrations.** The commentary annotation process in [29] allows the experts to pause at any instant and provide their feedback. Thus, the clip $\mathcal{V}_1$ could have the start of a basketball jump shot, whereas the clip $\mathcal{V}_2$ could also have content before the shot and the follow-through. Thus, we temporally align the learner and the expert video in the collection $\mathcal{C}$ to ensure we are capturing the same step of a demonstration.

For all pairs of clips in the collection $\mathcal{C}$, we compare the two poses $P$ to quantify their alignment (similarity) using Procrustes-aligned Mean Per Joint Position Error (PA-MPJPE) [27], a translation-invariant and body shape-invariant measure. Note that the video component of the input (Sec. 4) is complementary, as it does capture the person's overall movement in the space.

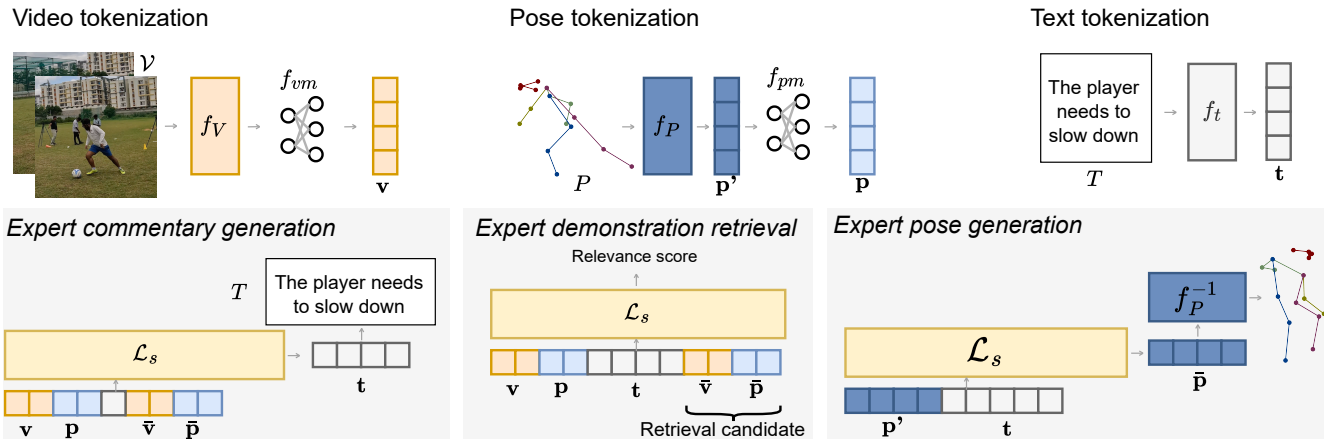


Figure 3. **Model overview.** We tokenize individual modalities using a modality-specific architecture (top). Once all the modalities are encoded as tokens, we use a large language model to learn expert commentary generation, demonstration retrieval, and pose generation. At inference, the model only takes the learner demonstration video $\mathcal{V}$. See text for details.

Denote the commentary timestamp as $t_1$ and $t_2$ for the two videos. We first choose a fixed window around $t_1$, say $[t_1^a, t_1^b]$ such that $t_1^a < t_1 < t_1^b$. Next, we find the corresponding time window $[t_2^a, t_2^b]$ with $t_2^a < t_2 < t_2^b$ in $\mathcal{V}_2$ such that the PA-MPJPE error is the minimum. Specifically,

$$[t_2^a, t_2^b] = \underset{[t_2^i, t_2^j] \in \mathcal{V}_2}{\text{argmin}} \text{ PA-MPJPE} \left( \mathcal{V}_1[t_1^a, t_1^b], \mathcal{V}_2[t_2^i, t_2^j] \right)$$
$$\text{such that } |t_2^j - t_2^i| = |t_1^b - t_1^a|.$$

We obtain a subset from $\mathcal{C}$ that contains the aligned video segments. There can be some video pairs where the demonstration can be very different and hence no appropriate match exists. Therefore, we only keep the top-k pairs with maximum alignment for every incorrect execution.

Overall, we obtain a novel dataset $\mathcal{D}$ that contains pairs of videos where the first video has an incorrect execution (on some body region) and the second video corrects it, thus obtaining the desired tuple $(\mathcal{V}, T, \bar{\mathcal{V}})$ for training and testing. See Fig. 2 (right) and examples in Supp. To ensure fair evaluation, we separately establish a clean gold standard test set, free of potential noise from LLM inference. The test set is manually verified (see Supp for details).

### 3.3. Architecture and training design

Next, we discuss the architecture that encodes the videos $\mathcal{V}$, poses $P$, and text $T$ and enables training the auto-regressive function $\mathcal{F}$. The overall idea is to encode all the representations into a text embedding space and use the strong capabilities of recent language models to obtain output text and pose tokens. This approach has been recently used in vision-conditioned language models [6, 35, 45, 55] for image and video captioning. This setup also allows for a unified architecture for various input and output combinations, as opposed to different input streams for individual modalities. Fig. 3 shows a schematic diagram of the architecture, and each part is explained below.

**Encoding video as tokens.** Each video $\mathcal{V}$ is an ego-exo clip pair.[1] The input thus allows the model to see both close-up hand-object interactions (more visible in the learner's ego view) as well as full-body poses in the scene context (more visible in the observer's exo view). We use a pre-trained video model to extract spatio-temporal features from both videos' frames, followed by a mapper to convert the video features to video tokens. Formally, $\mathbf{v} = f_{vm}(f_V(\mathcal{V}))$ where $f_V$ is a standard feature extractor (we use InternVideo2 [77]) and $f_{vm}$ is the visual mapper. See Fig. 3 (top left). The mapper is typically a low-parameter model that is trainable, whereas the high-parameter feature extractor is kept frozen. The mapper helps transform the visual mapper to be akin to text tokens—a popular strategy in visual instruction tuning models [35, 45].

**Encoding pose sequences as tokens.** To encode a pose sequence, we use a series of linear layers and MLP-mixer [25] to convert a single pose $P \in \mathbb{R}^{d \times 3}$ to an embedding that can be discretized using a codebook, i.e., $\mathbf{p}' = f_P(P)$. The architecture in [25] also contains a decoder that converts the embeddings back to human poses, i.e., $P = f_P^{-1}(\mathbf{p}')$, which we use in $\mathcal{F}_g$ to generate corrected poses. Similar to the above, we use a pose mapper to convert the tokens to embeddings. Formally, $\mathbf{p} = f_{pm}(f_P(P)) = f_{pm}(\mathbf{p}')$ is the pose token where $f_{pm}$ is the pose mapper. See Fig. 3 (top center). We concatenate embeddings from every frame to obtain the representation for the whole sequence. Having $f_{pm}$ allows training with fewer parameters and thus, we use $\mathbf{p}$ for learning $\mathcal{F}_t$ and $\mathcal{F}_r$. However, using $f_{pm}$ for generation $\mathcal{F}_g$ would require adding an inverting function for $f_{pm}$. Thus, we directly add $\mathbf{p}'$ to LLM tokens for pose generation $\mathcal{F}_g$. An innovative as-

[1] The exo view is the one annotated in Ego-Exo4D as having the maximum subject visibility.

pect of our approach is to encode pose as multimodal tokens and train with LLMs, unlike other work regressing the pose parameters from a special pose embedding [20] or using a dedicated pose transformer [49, 82].

**Encoding text as tokens.** Text tokenization is the standard process before inputting a text sentence to the LLM [59, 73], i.e., $\mathbf{t} = f_t(T)$ where $f_t$ is the tokenization function. See Fig. 3 (top right).

**Multi-modal sequence prediction.** The previous steps yield multimodal tokens $\mathbf{v}$, $\mathbf{p}$, and $\mathbf{t}$ for video, pose, and text, respectively. Next, we use the strong sequence prediction capabilities of large language models to obtain the desired output tokens based on the sequence of multi-modal inputs [6, 35, 45, 55].

For training $\mathcal{F}$ for expert commentary generation, we provide a sequence of learner video and pose tokens ($\mathbf{v}$ and $\mathbf{p}$) and the corrected pose tokens ($\bar{\mathbf{v}}$ and $\bar{\mathbf{p}}$ corresponding to $\bar{\mathcal{V}}$). We ask the model to predict the expert commentary token sequence. For the sequence prediction language model $\mathcal{L}_s$, we wish to obtain $\mathbf{t} = \mathcal{L}_s(\mathbf{v}, \mathbf{p}, \bar{\mathbf{v}}, \bar{\mathbf{p}})$. Consequently, the training objective is the standard cross-entropy loss, $\min_\theta \{-\log(\mathbf{t} \mid \mathbf{v}, \mathbf{p}, \bar{\mathbf{v}}, \bar{\mathbf{p}}; \theta)\}$, where $\theta$ are the parameters of the model. See Fig. 3 (bottom left). For consistency with the training of these language models, the sequence is formatted to be conversational in nature, e.g., "*provide an expert's commentary based on this pose sequence:*". Similarly, for expert demonstration retrieval, we wish to obtain a retrieval candidate $\bar{\mathbf{p}} = \mathcal{L}_s(\mathbf{v}, \mathbf{p}, \mathbf{t})$. Likewise, the training objective with parameters $\gamma$ is $\min_\gamma \{-\log(\bar{\mathbf{p}} \mid \mathbf{v}, \mathbf{p}, \mathbf{t}; \gamma)\}$, which we also use as the relevance score for retrieval during inference. Expert pose generation is trained with the same objective, except we use $\mathbf{p}'$ instead of $\mathbf{p}$ (and obtain $\bar{\mathbf{p}}'$). See Fig. 3 (bottom). Generating pose further requires converting back the pose tokens $\bar{\mathbf{p}}'$ to 3D joints using the pose decoder $f_P^{-1}$.

At inference time, as introduced in Sec. 3.1, we only use $\mathcal{V}$ as the input: $\mathcal{F}_t$ drops the expert demonstration and predicts the expert commentary $\hat{T}$. We use this predicted commentary as input for $\mathcal{F}_r$ and $\mathcal{F}_g$.

### 3.4. Implementation details

**Dataset and statistics.** Ego-Exo4D [29] contains 5,035 videos of participants doing activities across eight scenarios. We focus on three *physical* scenarios—basketball, soccer, and rock climbing—though our model is in principle generalizable to other physical skills, without any task-specific design, unlike [63, 78]. We use physical scenarios since the coaching feedback is groundable in the body regions, as opposed to procedural tasks like cooking, where the suggestions can be alternate ingredients or steps that are not visually present. Details and statistics are in Supp. Following the dataset creation process outlined in Sec. 3.2, we obtain a dataset of 25,505 training and 1,272 testing tuples of $(\mathcal{V}, T, \bar{\mathcal{V}})$. We choose $k = 5$ (train) and $k = 1$ (test) for min-k choice of correct demonstrations per incorrect execution. To reiterate, we manually examine the test set for correct commentary summary and body region labeling.

**Network architecture.** The video model $f_V$ is an InternVideo2 [77] encoder that provides strong visual representations. The pose encoder $f_P$ and decoder $f_P^{-1}$ are adapted from PCT [25], which learns compositional tokens from human poses. Since the original training in [25] uses 2D poses, we adapt it to use 3-dimensions and retrain on the human poses in [29]. Both the visual and pose mappers are low-parameter MLP layers [44]. Finally, we use Llama 3-8B [2] as the LLM multimodal encoder for sequence prediction. We finetune $\mathcal{F}_t$ and $\mathcal{F}_r$ for 5 epochs with a learning rate of $5 \times 10^{-5}$. Next, we modify the token dimension in $\mathcal{F}_g$ to accommodate pose tokens $\mathbf{p}'$, and hence, we fine-tune $\mathcal{L}_s$ when learning $\mathcal{F}_g$ with a learning rate of $5 \times 10^{-6}$ for 5 epochs. Video ($f_V$) and pose models ($f_P$, $f_P^{-1}$) are kept frozen. All the models are trained on 8 V100 32GB GPUs.

## 4. Experiments and results

We first discuss the baselines and ablations, followed by the evaluation setup and results for the three outputs—expert commentary generation ($\mathcal{F}_t$), expert demonstration retrieval ($\mathcal{F}_r$), and pose demonstration generation ($\mathcal{F}_g$). We also show qualitative examples and discuss the limitations.

**Baselines.** We compare with the following baselines:

- **InternVideo2-NN, InternVideo2-FT [77]**: Given a query video $\mathcal{V}$, the nearest neighbor baseline (NN) finds the most similar video by InternVideo2 [77] feature similarity in the training data and returns the corresponding commentary or demonstration, for all tasks. The FT version finetunes the model to contrastively match the learner demonstration with the expert commentary/demonstration.
- **VideoChat2 [36, 37], LLaVA [45]**: In these methods, we prompt SOTA video and image captioning models to generate commentary for an input demonstration. We use the log-likelihood loss to find the retrieved expert demonstration. These baselines evaluate if existing video captioning methods can provide expert feedback or retrieve expert demonstrations. We use the ego and "best exo" frames, same as for our method. Note that neither of these baselines are applicable for pose generation $\mathcal{F}_g$.
- **LLaVA-FT [45], LLaVA-FT w/ pose [45]:** These baselines are based on the SOTA visual-language method LLaVA [45] but trained on our dataset with the same text model, i.e., Llama 3-8B [2], for an apples-to-apples comparison. The "w/ pose" variant also takes the 3D pose coordinates in text form as input.
- **PoseScript [16], PoseFix [17]**: These two works enable pose-to-text and text-to-pose reasoning. The text generated or used for pose generation contains detail about the

| Method | Commentary Gen. B@4 | M | R-L | Demo Ret. R | medR ↓ | Pose Gen. P ↓ |
|---|---|---|---|---|---|---|
| InternVideo2-NN [77] | 42.1 | 46.9 | 49.3 | 13.5 | 198 | 161 |
| InternVideo2-FT [77] | 42.9 | 47.6 | 50.0 | 14.1 | 190 | 157 |
| VideoChat2 [37] | 27.8 | 44.3 | 41.9 | 14.9 | 183 | — |
| LLaVA [45] | 28.5 | 44.1 | 44.2 | 15.0 | 183 | — |
| LLaVA-FT [45] | 43.5 | 48.5 | 51.5 | 17.8 | 177 | — |
| LLaVA-FT w/ pose [45] | 43.6 | 48.5 | 51.7 | 18.0 | 172 | 150 |
| PoseScript/Fix [16, 17] | 24.1 | 44.5 | 46.3 | 15.9 | 182 | 182 |
| **ExpertAF** | **44.9** | **49.6** | **54.6** | **19.1** | **158** | **135** |
| **ExpertAF** w/o video | 44.6 | 49.4 | 54.2 | 18.7 | 161 | 139 |
| **ExpertAF** w/o pose | 44.2 | 49.3 | 54.2 | 18.6 | 163 | — |
| w/o alignment | 42.0 | 48.6 | 51.5 | 16.9 | 180 | 153 |
| w/ global pose | 43.0 | 47.9 | 52.6 | 16.5 | 184 | 150 |
| **ExpertAF** w/ full-sup | 45.8 | 50.9 | 55.7 | 22.5 | 146 | 131 |

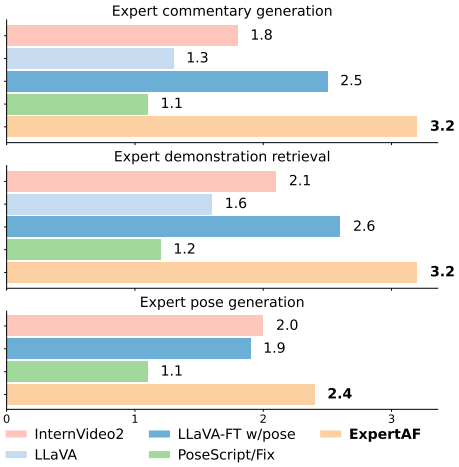


Table 1. **Results on automatic metrics (left) and human evaluation (right).** We break down results for the three outputs—expert commentary generation, expert demonstration retrieval, and expert pose generation. Our method outperforms all baselines and prior work on all tasks. The last row "w/ full-sup" uses privileged input (the demo video $\bar{\mathcal{V}}$) at inference. (B@4: BLEU-4, M: Meteor, R-L: ROUGE-L F1, R: recall@50, medR: median rank, P: PA-MPJPE). For all metrics higher is better, except medR and PA-MPJPE (↓). Our method is also rated higher by human raters on a Likert scale (min:1, max:4), compared to all the other methods (right). See text for details.

location of body parts, e.g., *the hands are raised*, as opposed to expert commentary. Hence, they let us evaluate if pose description is adequate for expert feedback. For pose generation, we evaluate if providing an expert commentary helps generate the desired expert demonstration. Both methods use SMPL [47] pose and hence we convert 3D pose to SMPL and vice versa [39], as needed.

**Ablations.** In addition to the strong baselines, we also compare the performance against ablations and variants of our design choices. **ExpertAF w/o pose** and **ExpertAF w/o video** evaluate the performance when only one modality is used. **ExpertAF w/o alignment** quantifies the need for temporal alignment (Sec. 3.2), while **ExpertAF w/ global pose** evaluates if just providing the model with one correct pose per activity chosen based on expert commentary is enough (vs. finding the closest correct pose sequence). **ExpertAF w/ full-sup** is a stronger variant with privileged input for inference—$\mathcal{F}_t$ uses $\mathcal{V}, \bar{\mathcal{V}}$ to predict $T$ and $\mathcal{F}_r$ and $\mathcal{F}_p$ uses both $\mathcal{V}$ and $T$. See Supp. for additional ablations of the choice of LLM $\mathcal{L}_s$, contribution of ego and exo videos, and joint training.

**Expert commentary generation.** To evaluate the text commentary, we use standard metrics BLEU-4 [60], METEOR [7] and ROUGE-L F1 [65], following prior work in evaluating text generation [14, 58] (higher is better).

Tab. 1 (left) shows the results. Our method outperforms all methods on all captioning metrics, with gains up to 3% over the best baseline trained on the same dataset. All the gains are statistically significant using paired t-test with $p < 0.05$. Fig. 4 (top) shows commentary generation by various methods. Firstly, SOTA methods LLaVA [45] and VideoChat2 [37], being captioning methods, yield a verbose description of the activity, rather than a critique. Fine-tuning LLaVA with our method improves the performance, but still it misses our use of expert demonstrations and pose sequences. Even using pose information as text only marginally improves "LLaVA-FT w/ pose", showing the effectiveness of our explicit pose encoding. Moreover, PoseScript [16] generates pose descriptions like *"head is turned to the left"* which is inadequate for feedback. Our results over the baselines shows the advantage of temporal alignment and a diverse expert demonstration set.

Our stronger variant "w/ full-sup" achieves an even better result by taking $\mathcal{V}$ and $\bar{\mathcal{V}}$ as inputs during inference. It is useful in cases where a learner has access to the expert demonstration but wants actionable feedback in language-form. Finally, Fig. 4 (bottom row) shows a failure case of commentary generation, showcasing the difficulty in pinpointing the exact mistake.

**Expert demonstration retrieval.** For retrieval, we use standard retrieval metrics: recall@k and median rank. We set k=50 to account for the fact that multiple expert demonstrations could suitably correct the given learner demonstration. The retrieval set contains $1,272$ samples, the same as the test set size. Higher recall and lower ranks are better.

Tab. 1 (middle) shows the results. Our method improves the median rank by 14 positions compared to the best baseline and a significantly better recall@50. Our method and the baselines show a similar trend as with the expert commentary generation above, the stronger baseline w/ "full-sup" (>4% better recall and 26 rank improvement) being crucial in cases where a learner has the commentary but wants a video exemplar to learn from.

**Expert pose generation.** Tab. 1 (right column) shows the expert pose generation results, with error measured by PA-MPJPE [27] (mm). Our novel pose generation method

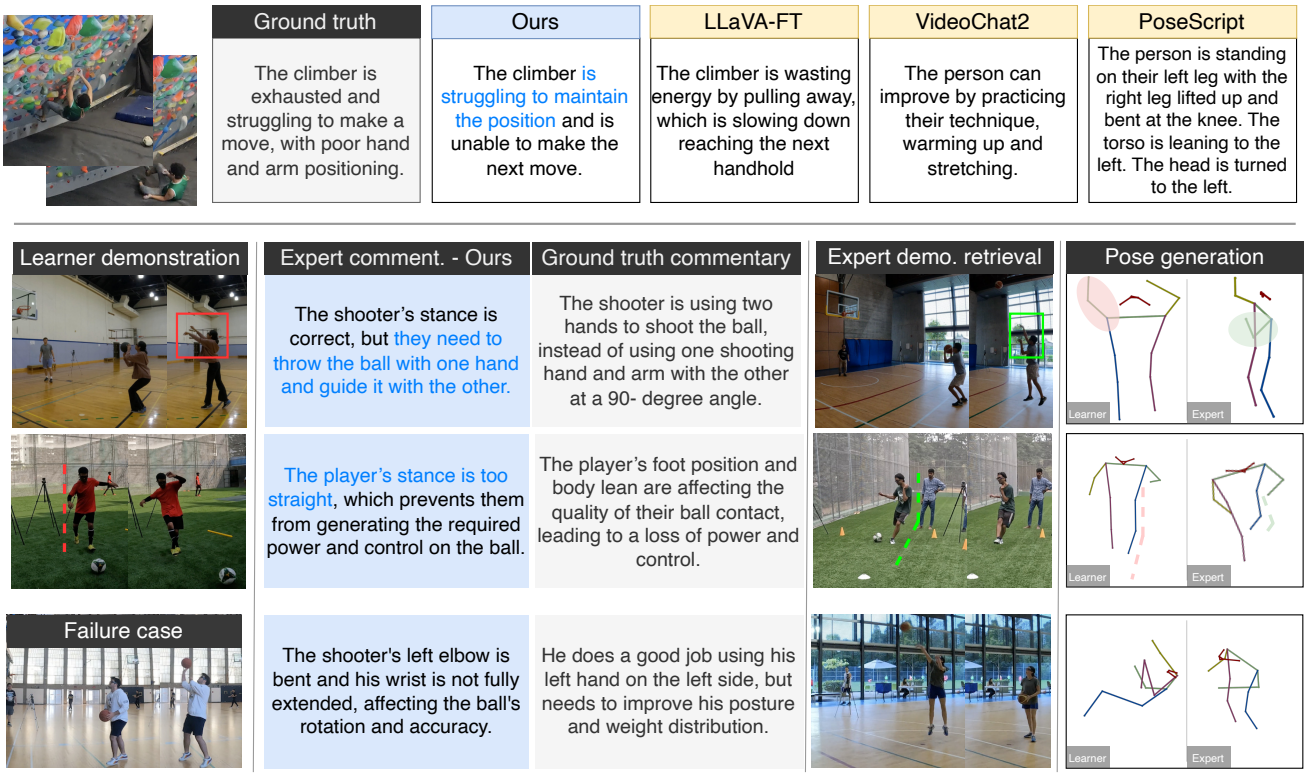


Figure 4. **Qualitative results.** (Top) Comparison of expert commentary generated by various baselines. (Second and third row) Examples of expert commentary generation, demonstration retrieval, and pose generation by our method. Notice the expert demonstration and pose generation corrects the mistake pointed out in the commentary, i.e., one hand is used to throw and the other to guide, and body position is improved to control better control and power. Colored text and marks (red for mistake, green for correction) are shown only for visualization. (Bottom) Failure cases. **See Supp. for video results.**

performs better than learning 3D position using text in "LLaVA-FT w/ pose". Furthermore, PoseFix [17] is designed to generate SMPL parameters based on a coarse modification text description. Our use of generation to output poses (vs. retrieval) is advantageous when the candidate retrieval set does not contain the correct demonstration; see Supp. See Fig. 4 and Supp. for qualitative examples including failure cases. The 3D pose sequences in the dataset are often auto-generated and have some noisy samples, discussed in Supp. Across the three tasks, our method is superior to its ablations.

**Human evaluation.** For all the tasks, we also solicit human evaluation (Tab. 1 right) to rate the quality of the generated text descriptions. Five raters (per scenario) uninvolved with this project rate the quality of each generated output by scoring on a Likert scale from 1 to 4 (higher is better);details in Supp. We ensure that the raters have a basic knowledge of the scenario they are rating (basketball, soccer, rock climbing).

Across all three tasks, human raters score our method the highest. A significant gap of up to 0.7 over the possible range $(1-4)$, as well as gains up to $3\times$, showcase ExpertAF's expert feedback quality. That said, there is naturally room for improvement on this new task. We find that while humans say our model excels on cases where the feedback is visually groundable, e.g. *incorrect hand angle*, it tends to fall short when the feedback is not directly visible, e.g. *the climber looks fatigued*. The human evaluation complements the automatic metrics above, overcoming the limitation that a "ground truth" commentary or a visual demonstration may capture only one of multiple possible errors in the learner's demonstration [23, 68] (e.g., for a basketball shot, both the hand placement and the jump could be incorrect, but only one may be mentioned in the ground truth).

## 5. Conclusion

We proposed a novel task and method to generate expert commentary and demonstrations from a learner's video. We develop a weakly-supervised training approach and benchmark for this problem. Our novel method fusing multimodal inputs from learner and expert demonstrations together with expert commentary results in state-of-the-art performance in actionable feedback, and lays the groundwork for accessible, affordable, and actionable AI coaching applications in the future.

**Acknowledgement.** UT Austin is supported in part by the IFML NSF AI Institute. Thanks to Fu-Jen Chu, Jing Huang and Xitong Yang for help with the Exo-Ego4D [29] pose extraction pipeline.

## 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. 2

[2] AI@Meta. Llama 3 model card, 2024. 4, 6, 1

[3] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23066–23078, 2023. 2

[4] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. What you say is what you show: Visual narration detection in instructional videos. *arXiv preprint arXiv:2301.02307*, 2023. 2

[5] 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. 1, 2

[6] Kumar Ashutosh, Zihui Xue, Tushar Nagarajan, and Kristen Grauman. Detours for navigating instructional videos. In *CVPR*, 2024. 1, 2, 5, 6

[7] 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. 7

[8] Tobias Baumgartner and Stefanie Klatt. Monocular 3d human pose estimation for sports broadcasts using partial sports field registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5108–5117, 2023. 2

[9] 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

[10] Gedas Bertasius, Aaron Chan, and Jianbo Shi. Egocentric basketball motion planning from a single first-person image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[11] 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. 2

[12] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators, 2024. 2

[13] 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. 2

[14] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023. 7

[15] Xin Chen, Anqi Pang, Wei Yang, Yuexin Ma, Lan Xu, and Jingyi Yu. Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision*, 129:2846–2864, 2021. 2

[16] 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. 2, 6, 7

[17] 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. 2, 6, 7, 8

[18] Nikita Dvornik, Isma Hadji, Hai Pham, Dhaivat Bhatt, Brais Martinez, Afsaneh Fazly, and Allan D Jepson. Flow graph to video grounding for weakly-supervised multi-step localization. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 319–335. Springer, 2022. 2

[19] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. Posegpt: Chatting about 3d human pose. *arXiv preprint arXiv:2311.18836*, 2023. 2

[20] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Learning disentangled avatars with hybrid 3d representations. In *CVPR*, 2024. 2, 6

[21] 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. 2

[22] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. Red: Reinforced encoder-decoder networks for action anticipation. *arXiv preprint arXiv:1707.04818*, 2017. 2

[23] Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur

Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online, 2021. Association for Computational Linguistics. 8

[24] Adélaïde Genay, Anatole Lécuyer, and Martin Hachet. Being an avatar "for real": a survey on virtual embodiment in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5071–5090, 2021. 2

[25] Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, and Han Hu. Human pose as compositional tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 660–671, 2023. 2, 5, 6

[26] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13505–13515, 2021. 2

[27] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023. 4, 7

[28] Andrew S Gordon. Automated video assessment of human performance. In *Proceedings of AI-ED*, page 10, 1995. 1, 2

[29] 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 *CVPR*, 2024. 1, 2, 3, 4, 6, 9

[30] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2

[31] William Hoff and Tyrone Vincent. Analysis of head pose accuracy in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 6(4):319–334, 2000. 2

[32] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2

[33] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[34] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7718–7727, 2019. 2

[35] 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, 6

[36] 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. 6

[37] 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. 6, 7

[38] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 2

[39] Yong-Lu Li, Xiaoqian Wu, Xinpeng Liu, Yiming Dou, Yikun Ji, Junyi Zhang, Yixing Li, Jingru Tan, Xudong Lu, and Cewu Lu. From isolated islands to pangea: Unifying semantic space for human action understanding. *arXiv preprint arXiv:2304.00553*, 2023. 7

[40] 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 *NeurIPS*, 2022. 2

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3

[42] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 2

[43] Daochang Liu, Qiyue Li, Tingting Jiang, Yizhou Wang, Rulin Miao, Fei Shan, and Ziyu Li. Towards unified surgical skill assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9522–9531, 2021. 3

[44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 6

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2, 5, 6, 7

[46] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C Kot. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2684–2701, 2019. 2

[47] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), 2015. 7

[48] Sabine Losch, Eva Traut-Mattausch, Maximilian D Mühlberger, and Eva Jonas. Comparing the effectiveness of individual coaching, self-coaching, and group training: How leadership makes the difference. *Frontiers in psychology*, 7: 175595, 2016. 1

[49] Vongani Maluleke, Lea Müller, Jathushan Rajasegaran, Georgios Pavlakos, Shiry Ginosar, Angjoo Kanazawa, and Jitendra Malik. Synergy and synchrony in couple dances. *arXiv preprint arXiv:2409.04440*, 2024. 6

[50] Priyanka Mandikal and Kristen Grauman. Dexvip: Learning dexterous grasping with human hand pose priors from video. In *Conference on Robot Learning*, pages 651–661. PMLR, 2022. 2

[51] 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. 2

[52] Effrosyni Mavroudi, Triantafyllos Afouras, and Lorenzo Torresani. Learning to ground instructional articles in videos through narrations, 2023. 2

[53] 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. 1, 2

[54] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[55] Tushar Nagarajan and Lorenzo Torresani. Step differences in instructional video. In *CVPR*, 2024. 5, 6

[56] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022. 1

[57] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. *arXiv preprint arXiv:2207.09666*, 2022. 2

[58] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *European Conference on Computer Vision*, pages 167–184. Springer, 2022. 7

[59] 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 Welin-

der, 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

[60] 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. 7

[61] 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. 2

[62] Paritosh Parmar, Jaiden Reddy, and Brendan Morris. Piano skills assessment. In *2021 IEEE 23rd international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE, 2021. 3

[63] 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. 2, 6

[64] 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. 2

[65] Lin CY Rouge. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 7

[66] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[67] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14761–14771, 2023. 2

[68] Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. Large language models are not yet human-level evaluators for abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4215–4233, 2023. 8

[69] Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[70] 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. 2

[71] 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. 2

[72] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 2

[73] 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

[74] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[75] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[76] Tianyu Wang, Yijie Wang, and Mian Li. Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, pages 668–678. Springer, 2020. 3

[77] 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. 2, 5, 6, 7

[78] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(12):4578–4590, 2020. 1, 2, 6

[79] 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. 2

[80] 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. 2

[81] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 2

[82] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14730–14740, 2023. 6

[83] Zhe Zhang, Chunyu Wang, Weichao Qiu, Wenhu Qin, and Wenjun Zeng. Adafuse: Adaptive multiview fusion for accurate human pose estimation in the wild. *International Journal of Computer Vision*, 129:703–718, 2021. 2

[84] 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. 2

[85] 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. 2

[86] Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, and Yizhou Wang. Motionbert: A unified perspective on learning human motion representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15085–15099, 2023. 2

[87] 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. 2