This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.

Except for this watermark, it is identical to the accepted version;

the final published version of the proceedings is available on IEEE Xplore.

# **Pose-Aware Weakly-Supervised Action Segmentation**

Seth Z. Zhao<sup>12\*†</sup> I

**Reza Ghoddoosian<sup>1\*‡</sup> Isht Dwivedi<sup>1</sup>** <sup>1</sup>Honda Research Institute, USA

**Nakul Agarwal<sup>1</sup> Behzad Dariush<sup>1</sup>** <sup>2</sup>UC Berkeley

### Abstract

Understanding human behavior is an important problem in the pursuit of visual intelligence. A challenge in this endeavor is the extensive and costly effort required to accurately label action segments. To address this issue, we consider learning methods that demand minimal supervision for segmentation of human actions in long instructional videos. Specifically, we introduce a weakly-supervised framework that uniquely incorporates pose knowledge during training while omitting its use during inference, thereby distilling pose knowledge pertinent to each action component. We propose a pose-inspired contrastive loss as a part of the whole weakly-supervised framework which is trained to distinguish action boundaries more effectively. Our approach, validated through extensive experiments on representative datasets, outperforms previous state-of-the-art (SOTA) in segmenting long instructional videos under both online and offline settings. Additionally, we demonstrate the framework's adaptability to various segmentation backbones and pose extractors across different datasets.

## **1. Introduction**

Recognizing human actions in a long instructional videos holds immense significance in facilitating comprehension and learning for intelligent systems. By accurately identifying and understanding human actions depicted in videos, human-machine interaction systems can interpret the sequential steps involved in performing complex tasks. This comprehension aids in skill acquisition and contributes toward enhancing human centered intelligent systems, human performance evaluation, and monitoring in various industrial applications. One big challenge lies in the fact that frame-level labeling of these videos demands extensive and costly human labor. Thus, a significant amount of research is dedicated to understanding human actions in long videos with *minimal human-crafted supervision*.

In this paper we study weakly-supervised learning meth-

ods for human action segmentation in long instructional videos. In this context, we work with video sequences paired with an ordered list of action labels (transcript), but without knowing the specific start and end times for each action. The main focus is to incorporate pose information into a weakly supervised framework, emphasizing its crucial role in temporal segmentation of human actions in the absence of frame-level labels. Pose information is informative due to its ability to encode rich information about body movements, gestures, and interactions, which are essential for a nuanced understanding of human actions. Specifically, pose information enables the decomposition of each action into a series of more granular representations, leading to more discriminative features within the same action segment and aiding in the identification of similarities across different actions. For instance, in an assembly task, the action "fasten screw" can be broken down into reaching for the screw and rotating the screw, each characterized by unique poses despite sharing the same action label. Moreover, it has been shown [12] that pose information is particularly discriminative during action transitions, making it a powerful feature to supervise learning and estimate the start and end times of actions in videos with weak labels, where detailed frame-level annotations are absent.

The proposed method leverages both appearance and gesture cues by combining RGB and pose modalities. While extracting pose information can be beneficial, it also introduces significant computational cost that can hamper real-time performance in interactive applications. To address this issue, we propose a framework that infuses pose information from a standard pose estimator into the RGB frame encoder during training. However, at test time, our method relies exclusively on the RGB modality. In particular, we employ a contrastive learning objective [9, 10, 27, 45, 63] to enable the model to differentiate between corresponding and non-corresponding RGB and pose features. By establishing a frame-level correspondence between RGB and pose features to create a positive pair, and by introducing a pose-based technique to identify negative pairs, our method facilitates the learning of features within a combined pose-RGB space.

Incorporating pose not only provides supplementary

<sup>\*</sup>These authors contribute equally.

<sup>&</sup>lt;sup>†</sup>Work done as intern at Honda Research Institute, USA.

 $<sup>^{</sup>t}Contact: reza_ghoddoosian@honda-ri.com.$ 

supervision, but also seamlessly distills knowledge into the RGB encoder, eliminating the reliance of pose features during inference. Our experimental evaluations on ATA[22], IKEA ASM[1], and Desktop Assembly[31] datasets demonstrate versatility and enhanced performance across various segmentation frameworks. The approach remains robust irrespective of the pose extractor used and is effective in both online and offline settings. Here, "online" refers to causal inference in streaming videos for interactive applications, whereas "offline" refers to post-analysis of pre-recorded videos.

The contributions of this paper are summarized as follows:

- This work is the first to integrate pose information into a weakly-supervised action segmentation framework establishing a segmentation approach, where despite the reliance on pose data during training, the model performs inference using only the conventional RGB modality, as commonly referenced in the literature [22, 35, 41, 48].
- We introduce our pose-based contrastive learning loss to distill pose knowledge into the RGB encoder, enhancing its capability to detect action boundaries in weakly-labeled untrimmed videos. This is achieved by utilizing the raw pose similarity across different frames to identify negative pairs for our loss.
- Through comprehensive ablation studies and rigorous testing on a variety of video datasets, including ATA, IKEA ASM, and Desktop Assembly, we demonstrate the versatility and broad applicability of our approach. Our method not only results in performance improvement across different segmentation frameworks and pose extraction tools but also proves effective in both online and offline scenarios.

## 2. Related Works

## 2.1. Weakly-Supervised Action Segmentation

Training action segmentation models under the weak supervision of transcripts was mainly initiated by [2]. Since then, many others [8, 21, 30, 47, 48] have proposed iterative [15, 22, 35, 41, 47, 48] or end-to-end [7, 52] approaches to align video frames to a given sequence of actions during training. However, they only use RGB-based features (I3D<sup>1</sup>[5] or iDT[29]) as input during both "inference and training". More similar to us, [20] use multi camera view points, only in training, to estimate more accurate frame-level pseudo labels. Consequently, they can segment videos using single view point input at test time. In contrast to all previous methods, we are the first to utilize pose to guide training and instill skeleton knowledge to the stan-

dard RGB-based features for inference.

#### 2.2. Pose in Action Understanding

There has been extended research in exploiting pose information for various video understanding tasks. Many papers focus on skeleton-based action recognition [28, 32, 33, 39, 42, 62, 64, 65], detection [13, 16, 17], and anomaly detection [17]. In these works pose information is used as the sole input [16, 17, 32, 33, 39, 42, 64, 65] or combined with RGB frames [13, 28, 62] to classify actions. In addition, [49] and [54] have further utilized contrastive loss between text and pose representations for action recognition and anomaly detection, respectively, in short videos. Similar to us, [58] apply pose to action segmentation. Specifically, they improve a skeleton estimator using self-supervised generative models. However, unlike our framework, all the aforementioned methods use pose in both training and inference time.

Our work is more aligned with recognition and detection methods that distill cross-modal knowledge from optical flow [34, 44], pose [11, 12, 14, 46] or depth [18, 19] to RGB encoders during training, so that at test time no modality except RGB is required. However, these methods are trained on fully labeled videos. In particular, although [25] does not require annotations for distillation through a pose reconstruction loss, they assume that an approximate bounding box for the athlete is provided in each frame. Also, fully-labeled sports videos are used for training in [25] for action recognition. Meanwhile, we are the first to take advantage of pose to guide weakly-supervised training and segment long videos into fine grained actions. In the area of self supervised learning, [51] uses cross modal similarity/dissimilarity such that features corresponding to all frames of a segment lie close in the latent space. In contrast, our RGB encoder learns to breakdown such segments into more discriminative pose-based representations by distilling pose knowledge in training.

#### 2.3. Contrastive Learning in Video Understanding

Contrastive learning is a popular solution for learning strong representations among multimodal interactions in both pretraining and multi-tasking settings [23, 36, 37, 45, 53, 55, 59, 60]. At a high level, contrastive learning contrasts samples against each other to learn features that are common and different between labels. Introduced in [45], CLIP is pretrained using a constrastive loss function to learn image representations from text. [27] expands on the contrastive objective of CLIP by producing challenging negative captions for every image-caption pair and selecting robust alternative images. [63] proposed a triplet contrastive loss objective based on InfoNCE [9] to draw together the embeddings of corresponding image-text pairs, while simultaneously separating non-matching pairs. In the video understanding domains, contrastive learning is often used as a

<sup>&</sup>lt;sup>1</sup>For simplicity, without loss of generality, we consider I3D features RGB-based although more accurately they are a mix of RGB and optical flow[61] streams.

pretraining objective in [3, 6, 38, 56, 57]. These methods often facilitate the representation learning of videos by leveraging vast amount of language transcripts [56, 57] or simply transferring the knowledge learned from image-text alignments [38]. In our work, we aim to leverage contrastive learning methods to facilitate video representation learning with the help of pose features.

## 3. Method

In this section, we present the problem formulation and an overview of the proposed pipeline. We then detail the pose encoding process and elaborate on the proposed contrastive losses.

#### 3.1. Problem Formulation

We formulate our task of weakly-supervised video action segmentation as follows. Given a video  $\mathbf{x}_1^t = (x_1, ..., x_t)$ with t frames and a "single person", the goal of the segmentation model is to segment a test video into a sequence of n actions  $\mathbf{a}_1^n = (a_1, ..., a_n)$  and their duration  $\mathbf{l}_1^n = (l_1, ..., l_n)$ . Notice that in a weakly-supervised training setting, we are not given frame-level action labels and we could only assume a sequence of action labels (transcripts)  $\mathcal{T}_1^n = (\mathcal{T}_1, ..., \mathcal{T}_n)$  that occur throughout the video. During the inference stage, two modes of offline and online settings are employed following previous protocols [20, 22, 48]. In offline mode, the model processes the entire video before segmentation, whereas in online mode, the model segments in real-time, only accessing frames up to the current moment.

#### 3.2. Method Overview

Our approach leverages pose features for enhanced supervision, improving visual representations learning without requiring per-frame action labels. As depicted in Fig. 1, we input precomputed RGB features and human poses, extracted by any frozen off-the-shelf estimator, into our framework. These inputs are processed by individual shallow encoders and then mapped into a shared representation space with consistent feature dimensions. Subsequently, contrastive learning loss is applied to embeddings from both modalities, enabling the RGB encoder to learn semantically rich visual representations enriched by pose data during training. This RGB encoder is also shared with a chosen weakly-supervised segmentation framework to decode the final output. During training, in order to integrate our contrastive learning with the original segmentation task, we use a multi-task setting. Here, the model minimizes a joint optimization objective, allowing the RGB encoder to incorporate the pose data and steer the segmentation loss to identify correct action segments across the video. The resulting training loss is:

$$\mathcal{L}_{Final} = \mathcal{L}_{con} + \mathcal{L}_{segment},\tag{1}$$

where  $\mathcal{L}_{con}$  is our proposed pose-based contrastive loss and  $\mathcal{L}_{segment}$  is the segmentation loss adopted from any weakly-supervised segmentation baseline [20, 22, 48]. During inference, we solely employ the RGB encoder, omitting the pose stream entirely and making our pipeline generalizable to various baselines without impacting runtime performance.

### 3.3. Detailed Pipeline

In this section, the process of extracting pose embeddings is first explained, followed by how pose information is infused into the RGB embeddings through our contrastive learning method in untrimmed videos.

## 3.3.1. Pose Encoding

Given a frame at time t, raw pose  $p_t \in \mathbb{Z}^{K \times 2}$  is a collection of (x, y) coordinates for K human keypoints. Here, K represents the number of 2D keypoints extracted by an external pose extractor and  $\mathbb{Z}$  is the set of integers. Before inputting these raw keypoints to the pose encoder, we perform a normalization step to ensure they are unaffected by changes in perspective, rotation, and positional offset in the frame. Specifically, each keypoint is centered and scaled with respect to the "center of mass" of the human, which is determined by averaging the coordinates of all joints. Subsequently, we determine the angle required to rotate each adjusted keypoint so that the head and "center of mass" align vertically, sharing the same x coordinates. These normalized 2D keypoints,  $\overline{p}_t$ , are then fed into the pose encoder.

As shown in Eqs. 2-4, the encoder uses a light-weight two-layer MLP network to learn rich representations from the pose keypoints and map them to the joint RGB-pose space. Following the approach in [49], each encoder layer is structured with sequential steps of layer normalization, ReLU activation, and dropout, with a residual link between layers complemented by max-pooling and a linear projection function  $\Gamma$  to refine the dimensionality of the resultant pose embedding  $P_t$ . Further details of the pose normalization and encoder architecture can be found in the supplementary materials.

$$z_1 = \operatorname{dropout}(\operatorname{ReLU}(\operatorname{LayerNorm}(W_1 \overline{p}_t + b_1))), \quad (2)$$

$$z_2 = \operatorname{dropout}(\operatorname{ReLU}(\operatorname{LayerNorm}(W_2 z_1 + b_2))), \quad (3)$$

$$P_t = \Gamma(\operatorname{maxpool}(z_2 + z_1)). \tag{4}$$

#### 3.3.2. Pose-Supervised Contrastive Learning

In this paper, we aim to utilize contrastive loss to create a common space for embedding both pose and RGB data.



Figure 1. Framework overview: Pose information is used exclusively during training. During inference, only image input is considered, omitting the pose branch.

This guides the RGB encoder to learn and infuse pose information into its output embeddings from processing RGB frames alone. Given a video  $\mathbf{V}_1^T = (v_1, ..., v_T)$  with T frames, we extract frame-level pose embeddings  $\mathbf{P}_1^T$  =  $(P_1, ..., P_T)$ , and RGB embeddings  $\mathbf{I}_1^T = (I_1, ..., I_T)$ , where  $I_t$  is the output of the RGB encoder at frame t. Various networks, such as Transformers or CNNs, can be utilized to implement the RGB encoder. For each frame t, serving as the anchor frame in one modality (RGB or pose), we define  $\mathbb{A}(t)$  and  $\mathbb{A}(t)$  as the sets of positive and negative frames, respectively, from the other modality within the same video. These sets,  $\mathbb{A}(t)$  and  $\mathbb{A}(t)$ , help identify positive and negative instances from the alternate modality that form corresponding pairs with the anchor at frame t. Utilizing these pairings for the anchor frame t, the RGB to pose contrastive loss  $\mathcal{L}_{I2P}$  is designed to increase similarity in positive pairs and dissimilarity in negative pairs.

$$\mathcal{L}_{I2P} = -\frac{1}{T} \sum_{t \in [0,T)} \log \frac{\sum_{i \in \mathbb{A}(t)} exp(sim(I_t, P_i)/\tau)}{\sum_{j \in \{\mathbb{A}(t) \cup \overline{\mathbb{A}}(t)\}} exp(sim(I_t, P_j)/\tau)},$$
(5)

where  $\tau$  is the temperature parameter and *sim* denotes the similarity function. Similarly, we derive the pose to RGB contrastive loss  $\mathcal{L}_{P2I}$ :

$$\mathcal{L}_{P2I} = -\frac{1}{T} \sum_{t \in [0,T)} \log \frac{\sum_{i \in \mathbb{A}(t)} exp(sim(P_t, I_i)/\tau)}{\sum_{j \in \{\mathbb{A}(t) \cup \overline{\mathbb{A}}(t)\}} exp(sim(P_t, I_j)/\tau)}.$$
(6)

Our overall contrastive loss  $\mathcal{L}_{con}$  is defined as the sum of  $\mathcal{L}_{I2P}$  and  $\mathcal{L}_{P2I}$ . Since each video encompasses various actions without frame-level labels, identifying the set of positive and negative frames,  $\mathbb{A}(t)$  and  $\overline{\mathbb{A}}(t)$ , for the contrastive loss  $\mathcal{L}_{con}$  poses a challenge.

Given anchor frame at t, a vanilla method is matching in time across different modalities to create a positive pair, and consider any other frame from the other modality as a negative pair. Formally,  $\mathbb{A}(t) = \{t\}$  and  $\overline{\mathbb{A}}(t) = \{j | j \in [0, T) \land j \neq t\}$ .

While this vanilla contrastive learning is straightforward,



Figure 2. Two methods to mine positive and negative frames for a given anchor in our contrastive learning framework.

it suffers from identifying false negative pairs under two primary conditions. Firstly, in instances where a pose is held for an extended period within a specific action sequence. For example, consider the RGB embedding at frame t during the action "balance part" as the anchor. Due to the static nature of this action, many negative poses identified throughout this segment closely resemble the positive pose at time t. The second challenging scenario occurs when a similar pose is repeated across various actions and is incorrectly tagged as a negative pose simply because it appears at a different time than t. In the context of posesupervised contrastive learning, we maintain that the infused pose knowledge should not depend on the action labels, especially in the absence of ground-truth. Since pose configurations are building blocks of actions and can reoccur across different actions, it is important to transfer the pose knowledge into the RGB encoder without linking it to specific action categories. Consequently, we filter out any negative pairs that feature poses similar to the anchor's, regardless of their occurrence time (see Fig. 2). To achieve this, we introduce  $d_{t,j} = |\overline{p}_t - \overline{p}_j|$  as the distance between the normalized key points of frames t and j. Accordingly, we redefine the set of negative frames as  $\overline{\mathbb{A}}(t) = \{j | j \in [0, T) \land d_{t,j} \geq \delta\}$ , where  $\delta$  is a predefined threshold.  $\mathbb{A}(t) = \{t\}$  is the same as in the vanilla case.

## 4. Experiments

In this section, we describe our experiments, followed by ablation studies and results of our framework compared to various baselines on multiple datasets. In the end, we conclude with qualitative evidence of how pose contributes to more accurate temporal boundary detection. Additional experiments are included in the supplementary material.

#### 4.1. Experimental Settings

Datasets. We conduct our experiments on three publicly available instructional video datasets: the ATA dataset[22], the Desktop Assembly dataset[31], and the IKEA dataset[1]. The ATA dataset contains 1152 toy assembly videos, captured from four different view-points, with each video averaging 1.3 minutes long and 12.9 segments. It features 32 participants assembling three different toys with 15 action classes and 96 unique transcripts. We adhere to the standard subject-based splitting of this dataset for test and validation. The Desktop Assembly dataset contains 76 desktop assembly videos, amounting a total of 2 hours, annotated with 23 action classes and 6 similar transcripts. It is split into 59 training and 17 testing videos. Lastly, the IKEA dataset contains 1113 furniture assembly videos, recorded from three perspectives, with an average duration of 1.9 minutes. This dataset is categorized into 33 action classes and offers 5 different training/testing splits.

**Evaluation Metrics.** Following previous work [20, 22, 50], we use four metrics to evaluate our action segmentation performance. 1) *acc* represents the average frame-level accuracy. 2) *IoU* determines intersection-over-union ratio for each predicted segment, excluding the background frames. 3) *Edit* employs edit distance to assess the similarity between predicted and ground-truth transcripts. 4) F1@0.5 assesses the per-class F1 score for predicted segments with an IoU threshold of 0.5.

**Implementation Details.** To remain consistency with prior works, we extracted I3D [4] features from ATA and IKEA datasets, and for Desktop Assembly, we used ResNet [24] features. In experiments with DP[22] as the baseline, we modeled the video encoder with Transformers. For MuCon[52] and TASL[41] segmentation baselines, we used their existing temporal convolution and GRU network outputs for RGB embedding, respectively. For computational efficiency, pose keypoints were extracted every five frames by RTMPose Body2D [26].  $\delta = 0.15$  for the experiments on the Desktop dataset while for ATA,  $\delta$  is set to 0.05 and 0.2 for online and offline segmentation respectively. The effect of  $\delta$  is discussed in Section 4.3. All other parame-

Table 1. Main results on weakly-supervised online segmentation. Our proposed pose-inspire framework improves the previous methods across different datasets.

Dataset	Method	acc	IoU	Edit	F1@0.5
ATA[22]	Greedy[43]	60.2	53.5	47.8	41.4
	DP[22]	62.3	53.3	55.5	48.2
	DP + Ours	66.0	<b>58.7</b>	<b>56.9</b>	<b>51.2</b>
Desktop[31]	Greedy [43]	4.8	2.5	24.7	0.3
	DP[22]	10.5	5.1	36.8	2.3
	DP + Ours	<b>18.0</b>	<b>7.6</b>	<b>52.2</b>	<b>3.7</b>
IKEA[1]	Greedy [43]	53.0	27.0	41.5	23.6
	DP[22]	54.3	27.3	48.1	26.0
	DP + Ours	<b>54.4</b>	<b>27.7</b>	<b>48.4</b>	<b>26.2</b>

ters, such as the number of training iterations, are set as per baseline settings [22, 41, 52]. More implementation details are included in the supplementary material. We intend to release the code and all parameters upon acceptance.

#### 4.2. Comparison Results

In this section, we show our pose-supervised segmentation framework improves both online and offline results on multiple datasets and baselines.

Weakly-Supervised Online Segmentation. In Table 1, we demonstrate the impact of our pose-inspired contrastive loss in comparison with previous weakly-supervised online segmentation methods. Notice that during training, both Greedy and DP share the same network structure. However, at inference time, Greedy adopts a sliding window approach to predict per-frame actions while DP uses an unconstrained dynamic programming approach based on the available transcripts. As shown in Table 1, infusing pose information into the RGB encoder of DP elevates its performance across all four metrics and three datasets. Specifically, on ATA and Desktop Assembly Dataset, the IoU performance gain is about 5.7% and 2.1%, respectively. We associate the smaller improvements on the IKEA dataset mostly to its 5th split. In many videos of this split, the single person assumption is violated by background people, which negatively impacts our pose encoding accuracy. We provide split-wise results on the IKEA dataset in the supplementary material.

Weakly-Supervised Offline Segmentation. For the sake of completeness, we also integrate our pose-inspired contrastive framework into three state-of-the-art offline segmentation methods, i.e., DP[22], TASL[41], and MuCon[52]. As shown in Table 2, infusing pose can consistently improve their weakly-supervised performance on the ATA and Desktop Assembly datasets. Despite the differences in network architecture and segmentation technique among various methods, our framework can be adapted to all baselines without changing their original architecture.

Table 2. Main results on weakly-supervised offline segmentation. Our pose-inspired framework is integrated into different baselines, which results in performance improvement across different metrics and datasets. Results are based on the best  $\delta$  values.

Dataset	Method	acc	IoU	Edit	F1@0.5
ATA[22]	CDFL [35]	58.1	44.9	59.5	50.9
	MuCon [52]	46.4	33.5	53.7	32.2
	MuCon + Ours	48.3	32.3	54.5	33.5
	TASL [41]	39.3	27.5	55.7	27.5
	TASL + Ours	45.7	29.3	51.1	33.2
	DP [22]	65.1	55.7	65.5	59.3
	DP + Ours	68.5	61.0	69.5	63.8
Desktop[31]	CDFL [35]	16.5	10.7	81.9	7.2
	MuCon [52]	46.0	33.0	100.0	27.2
	MuCon + Ours	50.9	35.4	100.0	31.3
	TASL [41]	35.2	22.4	95.8	14.7
	TASL + Ours	41.2	27.1	96.6	20.7
	DP [22]	16.3	10.7	86.5	6.5
	DP + Ours	17.2	12.0	91.4	7.7

Table 3. Comparison of vanilla and pose-inspired contrastive learning in weakly-supervised segmentation in the Desktop dataset. Results with the DP and TASL baselines correspond to online and offline segmentation modes respectively.

Backbone	Method	acc	IoU	Edit	F1@0.5
DP [22]	Baseline	10.5	5.1	36.8	2.3
	$L_{con-vanilla}$	12.7	7.2	48.3	3.7
	$L_{con-pose}$	18.0	7.6	52.2	3.7
TASL [41]	Baseline	35.2	22.4	95.8	14.7
	$L_{con-vanilla}$	39.3	26.3	95.8	16.5
	$L_{con-pose}$	41.2	27.1	96.6	20.7

In particular, on Desktop Assembly videos, instilling pose knowledge into the MuCon encoder achieves new SOTA and improves *acc* and *F1* by up to approximately 5%. Also, the high *Edit* score on Desktop Assembly videos is due to the very similar 6 transcripts of this dataset. Conversely, DP stands out as the best baseline for ATA videos, owing to its design for segmenting unseen sequences in the ATA test set.

#### 4.3. Analysis and Ablation Studies

In this section, we first compare the performance of our proposed contrastive loss, then test our framework's robustness across various pose extractors, and finally examine the pose knowledge learned by the RGB encoder. We use DP [22] and TASL[41] as baselines for our ablation study in online and offline segmentation tasks respectively.

**Contrastive Learning Mining Techniques.** We compare the results of our proposed pose-based and vanilla mining techniques in Table 3 for both online and offline segmentation. In particular, we utilized DP [22] for online and TASL[41] for offline segmentation tasks. Both contrastive learning methods outperform the baseline in all weakly-supervised segmentation experiments, demonstrating how the RGB encoder significantly benefits from the pose knowledge infusion. In addition, Table 3 shows that vanilla learning is consistently inferior to the posesupervised method, as it introduces a higher number of false negative samples that confuse the RGB model. On the other hand, utilizing pose for mining negative and positive instances account for pose variations within the same segment. This leads to a more fine-grained understanding of human dynamics and improves the recognition of the class and time extent of each segment in long videos.

The sensitivity of the threshold  $\delta$  in the pose-supervised learning method is illustrated in Fig. 3. Notably, incorporating the pose-supervised loss enhances performance over the baseline across all threshold values.  $\delta$  varies from 0, where no negative frames are removed, to a sufficiently high value that leads to the removal of all negative samples for contrastive learning. As shown in Fig. 3, results converge to the baseline as all negative samples are removed. Also, Note that the vanilla approach is a special case of the posesupervised method when  $\delta = 0$ .

The statistics of the pose distance  $d_{t,j}$  between any two frames t and j of a video is sensitive to the view-point. Hence, in a dataset like ATA, which features multiple viewpoints, finding a fixed effective threshold across all views is challenging. This is because a threshold value that is low for one view may be too high for another, leading to the removal of true negative frames.

Pose Type Generalizability. We assess the robustness of our framework by examining its online and offline performances with three pose extractors. We employ RTMOPose extractor[40], RTMPose Body2D, and RTMPose Whole-Body2D extractors to integrate pose into our framework. The main difference between these pose extractors is the level of keypoint detail, as illustrated in Fig. 4. The RTM-Pose WholeBody2D extractor identifies 133 fine-grained keypoints across the face, hands, and body, whereas RTM-Pose Body2D and RTMOPose identify 17 sparser set of keypoints. As shown in Table 4, our framework outperforms DP and TASL baselines on both the ATA and Desktop Assembly datasets, regardless of the extractor used, indicating its adaptability to different levels of pose detail. Table 4 further suggests that a higher number of keypoints results in competitive or larger improvements, due to the more detailed pose representations. This improvement is more substantial in the Desktop Assembly dataset where fine-grained pose estimations are more accurate. Also, note that while RTMOPose[40] achieves faster inference speed, its performance on open-world human pose extraction is not as accurate as RTMPose[26] series.

**Pose Knowledge Transferability.** In this section, we explore the extent to which pose knowledge, acquired dur-



Figure 3. Sensitivity analysis of  $L_{con-pose}$  on DP[22] online segmentation and TASL[41] offline segmentation on both ATA[22] and Desktop[31] datasets. Note that x-axis represents threshold value and y-axis represents results of *acc* and *IoU*.



(a) ATA Dataset

(b) Desktop Assembly Dataset

Figure 4. Visualization of zero-shot pose extraction results on both ATA Dataset and Desktop Assembly Dataset. Note that first row and second row represent RTMPose Body2D (sparse keypoints) and RTMPose WholeBody2D (dense keypoints) results, respectively. Compared to the Body2D keypoints, Wholebody2D keypoints have 116 additional keypoints on hands and face.

Table 4. Results of our pose-supervised framework using RTMPose Body2D[26], RTMPose WholeBody2D[26], and RTMOPose[40] pose detectors. Note that for DP[22] and TASL[41] we use online and offline segmentation settings respectively. Our framework improves the performance compared to the baselines [22, 41] irrespective of the external pose extractor.

Dataset	Pose Extractor	acc	IoU	Edit	F1@0.5
ATA[22]	DP [22] (No Pose)	62.3	53.3	55.5	48.2
	with RTMPose Body2D	66.0	58.7	<b>56.9</b>	<b>51.2</b>
	with RTMPose WholeBody2D	64.6	57.5	55.8	50.6
	with RTMOPose	63.2	56.2	55.9	48.6
Desktop[31]	DP [22] (No Pose)	10.5	5.1	36.8	2.3
	with RTMPose Body2D	18.0	7.6	52.2	3.7
	with RTMPose WholeBody2D	<b>22.8</b>	<b>14.1</b>	<b>53.9</b>	<b>10.8</b>
	with RTMOPose	12.1	7.0	39.2	3.1
ATA[22]	TASL [41] (No Pose)	39.3	27.5	<b>55.7</b>	27.5
	with RTMPose Body2D	<b>45.7</b>	<b>29.3</b>	51.1	33.2
	with RTMPose WholeBody2D	44.9	28.2	55.1	<b>34.5</b>
	with RTMOPose	39.5	28.2	49.9	32.0
Desktop[31]	TASL [41] (No Pose)	35.2	22.4	95.8	14.7
	with RTMPose Body2D	41.2	27.1	<b>96.6</b>	20.7
	with RTMPose WholeBody2D	<b>41.6</b>	<b>28.0</b>	95.8	<b>23.3</b>
	with RTMOPose	36.3	24.6	96.0	18.7

ing training, is applied during inference in the absence of an explicit pose modality. To investigate this, we conduct an experiment where, rather than infusing pose knowledge, we extract pose keypoints during both training and inference phases. In this setup, pose and RGB embeddings are merged prior to input into the segmentation model, and trained with the same loss as our proposed method to allow for a direct comparison. The concatenation baseline serves as the upper bound of our proposed method. Table 5 shows the RGB encoder in our method effectively assimilates pose knowledge through contrastive learning, often yielding performance comparable to its upper bound, even without direct use of pose information during inference.

Additionally for more insight, in Table 6, we compare our pose to RGB distillation result to that of RGB to pose distillation as well as pose and RGB only segmentation results. Table 6 shows that pose features are less discriminative than RGB features for action segmentation. While the performance of pose-only inference (row 1) is improved upon RGB distillation (row 2), it can not still compete with even the RGB-alone baseline (row 3).

### 4.4. Qualitative Analysis

Fig. 5 shows that enabling the RGB encoder to understand human poses enhances the accuracy of segmentation models. This improvement is attributed to the model's ability to learn nuances of human poses and their variations within



Figure 5. Qualitative results of our pose-based contrastive learning in online (left) and offline (right) segmentation. Understanding finegrained human pose results in more accurate detection of action boundaries at test time.

Table 5. Comparison of pose knowledge distillation (no pose at inference) with the oracle (DP + Concat) where pose is used at inference too on weakly-supervised online segmentation.

Dataset	Method	acc	IoU	Edit	F1@0.5
ATA[22]	DP[22] (no pose in training and inference)	62.3	53.3	55.5	48.2
	DP + Ours (no pose in inference)	66.0	58.7	56.9	51.2
	DP + Concat (with pose in inference)	67.5	60.2	<b>58.1</b>	<b>55.1</b>
Desktop[31]	DP[22] (no pose in training and inference)	10.5	5.1	36.8	2.3
	DP + Ours (no pose in inference)	18.0	7.6	52.2	3.7
	DP + Concat (with pose in inference)	<b>19.1</b>	<b>9.3</b>	52.2	<b>8.3</b>

Table 6. Offline weakly-supervised segmentation results on the Desktop dataset with TASL as the baseline.

Training	Testing	acc	IoU	Edit	F1@0.5
Pose	Pose	28.7	18.2	91.6	11.3
RGB+Pose	Pose	30.2	19.5	93.4	12.6
RGB	RGB	35.2	22.4	95.8	14.7
RGB+Pose	RGB	41.2	27.1	96.6	20.7

the same segment or during transitions from one action to another. Effectively, the detection of action boundaries becomes more precise. For example, in the ATA dataset, frames at time  $T_1$  and  $T_2$  show that the action "take part" consists of two main poses: extending a hand to grasp the part and placing it on the block. The baseline fails to identify the first pose as part of the action "take part", whereas our method has learned the extension pose precedes the placing pose as part of a single action. This pattern is consistently observed three times in Fig. 5 (left).

Additionally, note the fine granularity of poses that the RGB encoder can capture in videos from Desktop Assembly. Particularly, our method is able to recognize the transition from "pickup HDD" at time  $T_2$  to "install HDD" at time  $T_3$ . Also, our method is able to recognize that the combination of the two poses at time  $T_4$  and time  $T_5$  correspond

to the action "connect wire". In this instance, the absence of pose knowledge in the baseline results in incorrect detection of an action transition. It is remarkable that that such detailed pose understanding is obtained without explicitly utilizing the pose modality during inference. Yet, there are still instances where the infused pose information is not sufficiently discriminative to accurately identify the correct action. For example, at time  $T_1$  the action "tighten screw" is incorrectly classified as "place fan".

# 5. Limitations

The proposed paper sheds light on the impact of off-theshelf pose estimation in weakly-supervised action segmentation. However, it suffers from two main limitations. Firstly, the choice of  $\delta$  is dependent on the viewpoint. Because the relative distance between joints vary across different viewpoints, finding an optimal value that works best from different viewpoints can be challenging. Secondly, our method is devised for single-person action segmentation, so in videos with background people, e.g. IKEA dataset, it requires additional heuristics to eliminate background poses.

## 6. Conclusion

We introduce a weakly supervised action segmentation framework that leverages human pose knowledge in long instructional videos. The framework explores interactions between video sequences and human pose sequences during training and avoids using pose features at inference. Extensive experiments demonstrate the efficacy of the method as it outperforms the previous SOTA in segmenting long instructional videos under both online and offline settings. Furthermore, our framework can be extended to various segmentation backbones, pose extractors, causal and non causal settings for several representative datasets.

## References

- [1] Yizhak Ben-Shabat, Xin Yu, Fatemeh Saleh, Dylan Campbell, Cristian Rodriguez-Opazo, Hongdong Li, and Stephen Gould. The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 847–859, 2021. 2, 5
- [2] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 2
- [3] Digbalay Bose, Rajat Hebbar, Krishna Somandepalli, Haoyang Zhang, Yin Cui, Kree Cole-McLaughlin, Huisheng Wang, and Shrikanth Narayanan. Movieclip: Visual scene recognition in movies. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2083–2092, 2023. 3
- [4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017. 5
- [5] 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. 2
- [6] Santiago Castro and Fabian Caba. FitCLIP: Refining largescale pretrained image-text models for zero-shot video understanding tasks. In 33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022. BMVA Press, 2022. 3
- [7] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019. 2
- [8] Xiaobin Chang, Frederick Tung, and Greg Mori. Learning discriminative prototypes with dynamic time warping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 8395–8404, 2021. 2
- [9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020. 1, 2
- [10] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020. 1
- [11] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. Mars: Motion-augmented rgb stream for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7882–7891, 2019. 2
- [12] Rui Dai, Srijan Das, and François Bremond. Learning an augmented rgb representation with cross-modal knowledge distillation for action detection. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 13053–13064, 2021. 1, 2

- [13] Rui Dai, Srijan Das, Saurav Sharma, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 45(2):2533– 2550, 2022. 2
- [14] Srijan Das, Rui Dai, Di Yang, and Francois Bremond. Vpn++: Rethinking video-pose embeddings for understanding activities of daily living. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9703–9717, 2021. 2
- [15] Li Ding and Chenliang Xu. Weakly-supervised action segmentation with iterative soft boundary assignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 6508–6516, 2018. 2
- [16] Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, and Alessandro Bergamo. Skeletr: Towards skeleton-based action recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13634–13644, 2023. 2
- [17] Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely Di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, and Fabio Galasso. Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2023.
- [18] Nuno C Garcia, Pietro Morerio, and Vittorio Murino. Modality distillation with multiple stream networks for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 103–118, 2018. 2
- [19] Nuno Cruz Garcia, Sarah Adel Bargal, Vitaly Ablavsky, Pietro Morerio, Vittorio Murino, and Stan Sclaroff. Distillation multiple choice learning for multimodal action recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2755–2764, 2021. 2
- [20] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, Chiho Choi, and Behzad Dariush. Weakly-supervised online action segmentation in multi-view instructional videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13780–13790, 2022. 2, 3, 5
- [21] Reza Ghoddoosian, Saif Sayed, and Vassilis Athitsos. Hierarchical modeling for task recognition and action segmentation in weakly-labeled instructional videos. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pages 1922–1932, 2022. 2
- [22] Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, and Behzad Dariush. Weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10128–10138, 2023. 2, 3, 5, 6, 7, 8
- [23] Haoming Guo, Seth Z. Zhao, Jiachen Lian, Gopala Anumanchipalli, and Gerald Friedland. Enhancing gan-based vocoders with contrastive learning under data-limited con-

dition. In 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW), pages 480–484, 2024. 2

- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 5
- [25] James Hong, Matthew Fisher, Michaël Gharbi, and Kayvon Fatahalian. Video pose distillation for few-shot, fine-grained sports action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9254– 9263, 2021. 2
- [26] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose, 2023. 5, 6, 7
- [27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. arXiv preprint arXiv:2004.11362, 2020. 1, 2
- [28] Sangwon Kim, Dasom Ahn, and Byoung Chul Ko. Crossmodal learning with 3d deformable attention for action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10265–10275, 2023.
- [29] Hilde Kuehne, Juergen Gall, and Thomas Serre. An end-toend generative framework for video segmentation and recognition. In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pages 1–8. IEEE, 2016. 2
- [30] Hilde Kuehne, Alexander Richard, and Juergen Gall. Weakly supervised learning of actions from transcripts. *Computer Vision and Image Understanding*, 163:78–89, 2017. 2
- [31] Sateesh Kumar, Sanjay Haresh, Awais Ahmed, Andrey Konin, M Zeeshan Zia, and Quoc-Huy Tran. Unsupervised action segmentation by joint representation learning and online clustering. In *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition, pages 20174– 20185, 2022. 2, 5, 6, 7, 8
- [32] Jungho Lee, Minhyeok Lee, Suhwan Cho, Sungmin Woo, Sungjun Jang, and Sangyoun Lee. Leveraging spatiotemporal dependency for skeleton-based action recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10255–10264, 2023. 2
- [33] Jungho Lee, Minhyeok Lee, Dogyoon Lee, and Sangyoun Lee. Hierarchically decomposed graph convolutional networks for skeleton-based action recognition. In *Proceedings* of the IEEE/CVF International Conference on Computer Vision, pages 10444–10453, 2023. 2
- [34] Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, and Hyeran Byun. Decomposed cross-modal distillation for rgb-based temporal action detection. In *Proceedings of* the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2373–2383, 2023. 2
- [35] Jun Li, Peng Lei, and Sinisa Todorovic. Weakly supervised energy-based learning for action segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6243–6251, 2019. 2, 6

- [36] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021. 2
- [37] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2
- [38] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen clip models are efficient video learners. arXiv preprint arXiv:2208.03550, 2022. 3
- [39] Xingyu Liu, Sanping Zhou, Le Wang, and Gang Hua. Parallel attention interaction network for few-shot skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1379–1388, 2023. 2
- [40] Peng Lu, Tao Jiang, Yining Li, Xiangtai Li, Kai Chen, and Wenming Yang. RTMO: Towards high-performance onestage real-time multi-person pose estimation, 2023. 6, 7
- [41] Zijia Lu and Ehsan Elhamifar. Weakly-supervised action segmentation and alignment via transcript-aware union-ofsubspaces learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8085–8095, 2021. 2, 5, 6, 7
- [42] Zhengzhi Lu, He Wang, Ziyi Chang, Guoan Yang, and Hubert PH Shum. Hard no-box adversarial attack on skeleton-based human action recognition with skeletonmotion-informed gradient. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4597– 4606, 2023. 2
- [43] Ran Xu Richard Socher Caiming Xiong Mingfei Gao, Yingbo Zhou. Woad: Weakly supervised online action detection in untrimmed videos. In CVPR, 2021. 5
- [44] Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, and Tinne Tuytelaars. Multimodal distillation for egocentric action recognition. In *Proceedings* of the IEEE/CVF International Conference on Computer Vision, pages 5213–5224, 2023. 2
- [45] 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. 1, 2
- [46] Dominick Reilly and Srijan Das. Just add?! pose induced video transformers for understanding activities of daily living. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18340–18350, 2024. 2
- [47] Alexander Richard, Hilde Kuehne, and Juergen Gall. Weakly supervised action learning with rnn based fine-to-coarse modeling. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 754–763, 2017.
   2
- [48] Alexander Richard, Hilde Kuehne, Ahsan Iqbal, and Juergen Gall. Neuralnetwork-viterbi: A framework for weakly

supervised video learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7386–7395, 2018. 2, 3

- [49] Fumiaki Sato, Ryo Hachiuma, and Taiki Sekii. Promptguided zero-shot anomaly action recognition using pretrained deep skeleton features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6471–6480, 2023. 2, 3
- [50] Saif Sayed, Reza Ghoddoosian, Bhaskar Trivedi, and Vassilis Athitsos. A new dataset and approach for timestamp supervised action segmentation using human object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3132–3141, 2023. 5
- [51] Anshul Shah, Benjamin Lundell, Harpreet Sawhney, and Rama Chellappa. Steps: Self-supervised key step extraction and localization from unlabeled procedural videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10375–10387, 2023. 2
- [52] Yaser Souri, Mohsen Fayyaz, Luca Minciullo, Gianpiero Francesca, and Juergen Gall. Fast weakly supervised action segmentation using mutual consistency. *IEEE Transactions* on Pattern Analysis and Machine Intelligence, 2021. 2, 5, 6
- [53] Kehan Wang, Seth Z. Zhao, David Chan, Avideh Zakhor, and John Canny. Multimodal semantic mismatch detection in social media posts. In *Proceedings of IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, 2022.
   2
- [54] Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, and Lei Zhang. Generative action description prompts for skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10276–10285, 2023. 2
- [55] Chenfeng Xu, Tian Li, Chen Tang, Lingfeng Sun, Kurt Keutzer, Masayoshi Tomizuka, Alireza Fathi, and Wei Zhan. Pretram: Self-supervised pre-training via connecting trajectory and map. arXiv preprint arXiv:2204.10435, 2022. 2
- [56] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. VLM: Task-agnostic videolanguage model pre-training for video understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4227–4239, Online, 2021. Association for Computational Linguistics. 3
- [57] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. VideoCLIP: Contrastive pretraining for zero-shot video-text understanding. In *Proceedings of the* 2021 Conference on Empirical Methods in Natural Lan-

*guage Processing (EMNLP)*, Online, 2021. Association for Computational Linguistics. 3

[58] Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, and Francois Bremond. Lac-latent action composition for skeleton-based action segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 13679–13690, 2023. 2

- [59] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. Vision-language pre-training with triple contrastive learning. 2022. 2
- [60] Yuan Yao, Qianyu Chen, Ao Zhang, Wei Ji, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. In *Proceedings of EMNLP*, 2022. 2
- [61] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-1 1 optical flow. In *Joint pattern recognition symposium*, pages 214–223. Springer, 2007. 2
- [62] Haosong Zhang, Mei Chee Leong, Liyuan Li, and Weisi Lin. Pgvt: Pose-guided video transformer for fine-grained action recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6645–6656, 2024. 2
- [63] Zhenyu Zhang, Yuming Zhao, Meng Chen, and Xiaodong He. Label anchored contrastive learning for language understanding. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1437–1449, Seattle, United States, 2022. Association for Computational Linguistics. 1, 2
- [64] Huanyu Zhou, Qingjie Liu, and Yunhong Wang. Learning discriminative representations for skeleton based action recognition. In *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition, pages 10608– 10617, 2023. 2
- [65] Yisheng Zhu, Hu Han, Zhengtao Yu, and Guangcan Liu. Modeling the relative visual tempo for self-supervised skeleton-based action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13913–13922, 2023. 2