PeVL: Pose-Enhanced Vision-Language Model for Fine-Grained Human Action Recognition

Haosong Zhang[1,2], Mei Chee Leong[1], Liyuan Li[1], Weisi Lin[2]
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore[1]
Nanyang Technological University, Singapore[2]
haosong001@e.ntu.edu.sg

Abstract

Recent progress in Vision-Language (VL) foundation models has revealed the great advantages of cross-modality learning. However, due to a large gap between vision and text, they might not be able to sufficiently utilize the benefits of cross-modality information. In the field of human action recognition, the additional pose modality may bridge the gap between vision and text to improve the effectiveness of cross-modality learning. In this paper, we propose a novel framework, called ***Pose-enhanced Vision-Language (PeVL)*** *model, to adapt the VL model with pose modality to learn effective knowledge of fine-grained human actions. Our PeVL model includes two novel components: an Unsymmetrical Cross-Modality Refinement (UCMR) block and a Semantic-Guided Multi-level Contrastive (SGMC) module. The UCMR block includes Pose-guided Visual Refinement (P2V-R) and Visual-enriched Pose Refinement (V2P-R) for effective cross-modality learning. The SGMC module includes Multi-level Contrastive Associations of vision-text and pose-text at both action and sub-action levels, and a Semantic-Guided Loss, enabling effective contrastive learning with text. Built upon a pre-trained VL foundation model, our model integrates trainable adapters and can be trained end-to-end. Our novel PeVL design over VL foundation model yields remarkable performance gains on four fine-grained human action recognition datasets, achieving a new SOTA with a significantly small number of FLOPs for low-cost re-training.*[1]

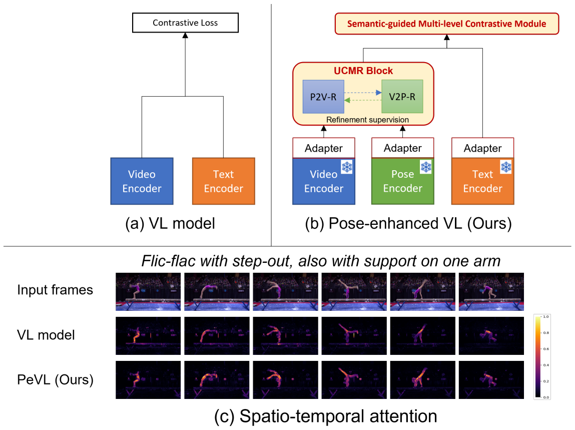


Figure 1. The top row shows the architecture comparison between (a) VL model and (b) our proposed PeVL, where the blocks with red texts in PeVL are the novel modules. The bottom row shows the effectiveness of PeVL cross-modality joint learning, where the images in the 3rd row shows much better spatial attentions on bodies and temporal attentions on sequential sub-actions learned by PeVL compared to existing VL model results.

1. Introduction

Understanding human actions through video is a complex and multifaceted challenge. Human beings perform the task by exploiting multiple intelligent capabilities, such as visual perception, body structure and motion recognition, and concept-level semantic descriptions (text). The majority of previous works focus on video modality utilizing convolutional networks or transformers. Recently, human action recognition has undergone substantial evolution due to the advent of Vision Language (VL) foundation models [21, 27, 41]. VL models have shown great advantages in cross-modality learning, aligning image and text features through shared embedding spaces derived from vast paired image-text datasets. The raw vision modality presents low-level detailed visual features, while the text provides high-level coarse semantic descriptions. While

[1] This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-GC-2022-005) and the Agency for Science, Technology and Research, under the AME Programmatic Funding Scheme A18A2b0046. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

prior efforts [6, 50, 54, 59] have fine-tuned VL models to improve video recognition, their efficacy in associating the vast disparity between visual and textual modalities remains a limitation. The additional pose modality (skeleton or joint points), which provides a latent semantic representation of the body structure and joint movement, might be helpful to bridge the gap.

Inspired by this motivation, in this paper, we propose a novel framework, named Pose-enhanced Vision-Language (PeVL) model, designed to achieve effective cross-modality learning among vision-pose-language (VPL) modalities with low-cost re-training. The overview of our architecture is illustrated in Figure 1. It contains two distinctively designed modules over the pre-trained VL foundation model. One module is the Unsymmetrical Cross-Modality Refinement (UCMR) block for visual-pose joint learning. It consists of two joint sub-blocks for: (i) pose-guided visual refinement which employs pose representation to guide the attention of visual learning, and (ii) vision-enriched pose refinement which introduces additional visual tokens to enhance the representation of pose tokens. Another module is the Semantic-Guided Multi-level Contrastive (SGMC) module to establish cross-modal associations at both coarse-grained (action) and fine-grained (sub-action) levels. The contrastive learning is then guided by the adaptive discrepancies between text and video representations. The effectiveness of spatial attention adaption on body and temporal attention alignment on sub-actions can be observed in the example images in Figure 1(c).

We conduct thorough evaluations on four benchmark datasets of fine-grained human action recognition. Our results demonstrate that PeVL is effective in bridging the gap between vision and language modalities, outperforming existing methods in achieving new SOTA. Our contributions are summarized as follows: (i) we present Pose-enhanced Vision-Language (PeVL), a novel framework that adapts VL foundation model with additional pose modality for fine-grained human action recognition. (ii) we propose UCMR block to achieve effective cross-modality learning between concise pose structural configurations and rich visual features. (iii) we introduce SGMC module for effective association between vision-text and pose-text for both action level generality and sub-action level alignment. (iv) extensive experimental results on fine-grained human action recognition benchmarks demonstrating improved performance over base models and outperforming SOTA.

2. Related Work

Action recognition Accurate representation of spatial and temporal information is pivotal for effectively recognizing actions in videos. Previous approaches to video action recognition involved a fusion of 2D or 3D convolutional layers and sequential models to capture spatial and temporal relationships [5, 16, 53]. More recently, researchers have introduced vision transformer-based architectures [1, 35, 57], which excel at modeling extensive spatio-temporal dependencies and have notably surpassed traditional convolutional counterparts. Another line of work [7, 13] exploits human pose as an additional modality for joint learning with video. PoseConv3D [13] extracts pose heatmaps and employs 3D CNN for joint video-pose learning, while ViLP [7] firstly integrates video, pose and text encoding utilizing 2D pose heatmaps and coarse-grained category names for joint embedding learning. Different from existing methods, we directly use the 2D body joint coordinates as pose input for learning structural temporal dynamics in fine-grained actions. We utilize video labels and text prompts that describes the atomic actions and action category to supervise feature alignment with video and pose modalities.

Adaption of Vision-language (VL) Models The emergence of VL models marked a significant turning point in the field, notably demonstrated by pioneering works such as CLIP [41] and ALIGN [21]. These groundbreaking models showcased the potential of large-scale pre-training on extensive datasets with abundant image-text pairs sourced from the web. To refine image-text alignment, these models adopt contrastive learning objectives. Previous techniques [8, 28, 47] often relied on object detectors to extract region features prior to pre-training, and later subsequent efforts introduced cross-attention layers and self-supervised learning objectives, encompassing tasks such as image-text matching [21, 24, 27, 58]. Recent progress, exemplified by works like ActionCLIP [51] and XCLIP [37], has embraced a multi-modal paradigm by extending the CLIP framework to encompass video comprehension. These innovative developments have revolved around adapting large-scale VL models for video understanding [6, 22, 42, 54]. This paper explores extending existing VL models by introducing an additional pose modality and multi-level supervised learning with text representation to enhance the understanding of fine-grained actions.

3. Our Method

The architecture of the proposed PeVL model is illustrated in Figure 2. It consists of three components: (1) three separated *unimodal encoders* with adapters for encoding video, pose and text inputs; (2) an *Unsymmetric Cross-Modality Refinement (UCMR) Block* for effective video-pose cross-modality learning; (3) a *Semantic-Guided Multi-level Contrastive (SGMC) Module* for both action and sub-action levels text-guided VPL joint learning. The technical details of the modules are described in Section 3.1 to Section 3.4.

3.1. Unimodal Encoders with Adapters

PeVL takes the raw video clip, the 2D pose body joints, and the texts as inputs. We use the CLIP image encoder

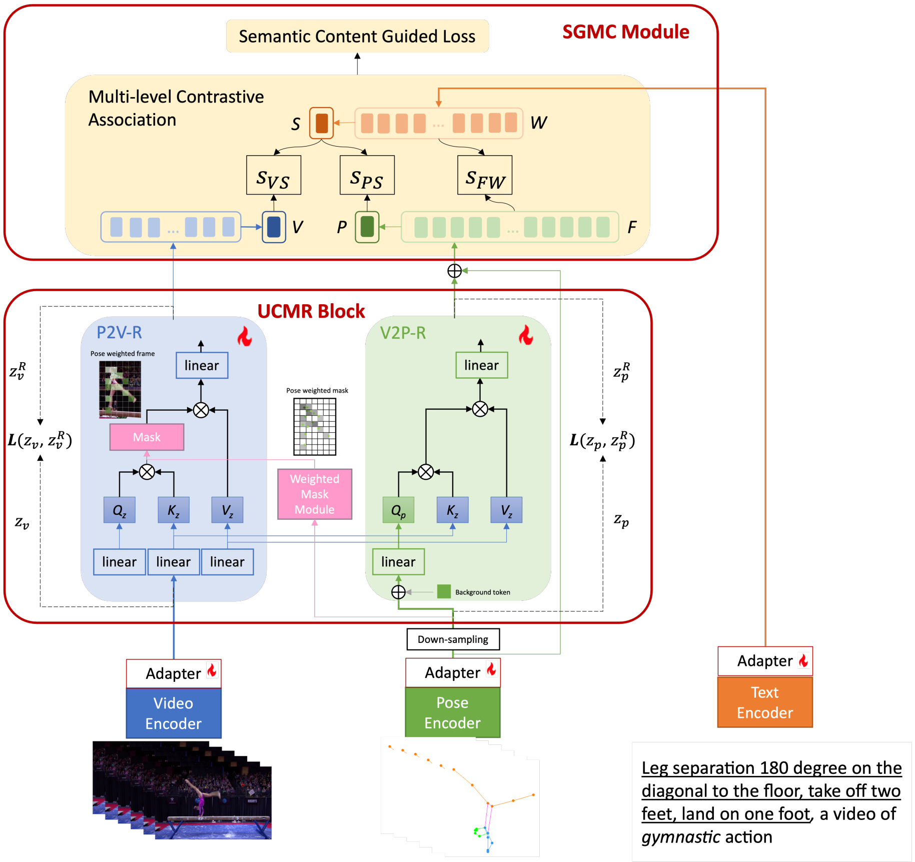


Figure 2. **PeVL Framework Architecture**. Our proposed framework consists of three components: (1) three *unimodal encoders* with adapters on VL foundation model for feature input; (2) a *Unsymmetric Cross-Modality Refinement Block* (UCMR Block) for video-pose cross-modality learning; (3) a *Semantic Guided Multi-level Contrastive Module* (SGMC Module) for text-supervised vision-text and pose-text contrastive learning.

for video and pose inputs and the CLIP text encoder for text inputs, with trainable adapters that have a bottleneck structure [59]. Our network benefits from large-scale pre-trained VL models, initializing with few new parameters for a strong starting point. We use "prompting" to adapt the video label sentence to the image-text pre-trained VL model. Text t is transformed into prompted textual input t' by appending prompts (e.g., "label, a video of action"). An optional coarse-grained action type prompt enriches text representations with prior knowledge. As an example, for the FineGym action in Figure 2, we add "gymnastic" to prompt, which will be "Leg separation 180 degree on the diagonal to the floor, take off two feet, land on one foot, a video of *gymnastic* action", which is text input in PeVL. We pass prompted text tokens through text encoder for words modeling to yield text representation z_t.

3.2. UCMR Block

Video and pose are unsymmetrical modalities containing rich low-level visual features and concise latent semantic representation of body structure respectively. To achieve effective cross-modality learning on video and pose, we design an *Unsymmetric Cross-Modality Refinement Block* with two distinct sub-blocks. The first one, named *Pose-adapted Vision Refinement (P2V-R)*, exploits the pose structure to guide the learning of video representation, reinforcing its attention on body joint regions in the image. In contrast, the second one, named *Video-enriched Pose Refinement (V2P-R)*, employs the related visual information from the video to enrich the pose representation, facilitating the learning of appearance-aware pose representation.

3.2.1 Pose-adapted Vision Attention (P2V)

The raw image has rich visual information. Hence, the learning algorithm simply on an image might not be able to effectively focus its attention on body parts. We construct a Weighted Mask Module (Figure 3(a)) for each video frame to guide the learning attention to the regions around pose joints. From a 2D pose detector, we obtain a set of

$\mathcal{J} \in \mathbb{R}^{T \times N_p \times 2}$ that denotes the (x, y) coordinates of the N_p body joints, for T number of pose frames. We then define a weighted pose mask $m \in \mathbb{R}^{H \times W}$ for each frame with spatial resolution $H \times W$, obtained by a normalised sum of weights of all body joints from the last layer of Pose Encoder at the $(x, y)^{th}$ pixel. We denote $m = 0$ if there is no body joint present at the $(x, y)^{th}$ pixel. $\mathcal{M} = \{m_i | i \in \{1, ..., T_v\}\}$ is a collection of weighted pose masks for all T_v frames, where $T_v < T$. To align with ViT inputs, $\mathcal{M}$ is decomposed into N_v disjoint patches, which corresponds to the patch size in video embedding z_v. Thus, each m in $\mathcal{M}$ indicates the weights of video patches that contain body joints within a frame. Essentially, P2V functions as local attention that modulates the video token representation containing weighted pose, as shown in Figure 3(a). P2V learns attention weights $\boldsymbol{\alpha}^{\text{P2V}}$ for a video token as:

$$\boldsymbol{\alpha}^{\text{P2V}} = \frac{\exp(\mathbf{Q_v} \mathbf{K_v}^{\intercal} \odot \boldsymbol{\mathcal{M}})}{\sum_{j=1}^{N_v} \exp(\mathbf{Q_v} \mathbf{K_v}_j^{\intercal} \odot \boldsymbol{\mathcal{M}}) + \mathbb{I}}, \qquad (1)$$

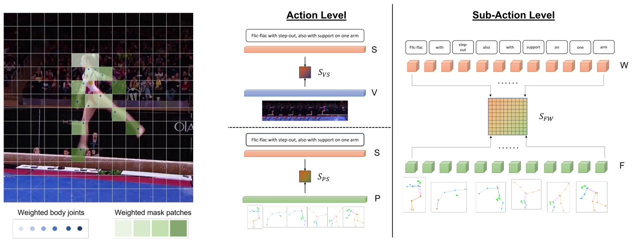
where indicator $\mathbb{I} = 1$ if $\mathcal{M} = 0$ else $\mathbb{I} = 0$, to solve the issue of zero denominator. Both $\mathbf{Q_v} \in \mathbb{R}^{N_v \times d}$ and $\mathbf{K_v} \in \mathbb{R}^{N_v \times d}$ are linear projected video tokens. Using $\mathcal{M}$, the spatial attention in P2V allows interaction among tokens containing weighted body joints in a single frame. The P2V is a plug-in module that can be inserted into the ViT architecture to induce learning of pose-guided representations. When inserted into a ViT, P2V will process tokens from the video encoder and return a set of video tokens enhanced with pose-adapted attention, namely refined video embeddings z_v^R.

3.2.2 Video-enriched Pose Representation (V2P)

Pose is represented by the spatial coordinates of joint points, where the visual context information of the human body has been lost. In V2P, we introduce visual tokens to enrich the pose tokens with related visual context information. It is designed by adopting multi-modal fusion methods [3, 12, 19, 55]. The pose tensor is first downsampled to a reduced number of frames (T_v) to match the indexing and temporal dimension of the video modality. We concatenate one additional token to the pose embedding to represent the background. Then, we group all pixels that belong to the corresponding pose tokens, including body joint tokens and a background token. V2P learns attention weights $\boldsymbol{\alpha}^{\text{V2P}}$ for a pose token as:

$$\boldsymbol{\alpha}^{\text{V2P}} = \frac{\exp(\mathbf{Q_p} \mathbf{K_v}^{\intercal})}{\sum_{j=1}^{N_v} \exp(\mathbf{Q_p} \mathbf{K_v}_j^{\intercal})}, \qquad (2)$$

where $\mathbf{Q_p} \in \mathbb{R}^{(N_p+1) \times d}$ is the linear projected pose tokens. The outputs of V2P can be perceived as refined pose embedding $z_p^{R'}$ based on video information. Afterwards, we combine $z_p^{R'}$ with the rest of the pose frames that have not been indexed, to form a combined pose representation z_p^R of T frames.



(a) Weighted mask module

(b) Multi-level Contrastive Association

Figure 3. (a) The image illustrates the weighted mask module in UCMR block for exploiting pose to guide the vision attention; (b) The illustration of Multi-level Contrastive Associations employed in SGMC block, which contains action-level video-text and pose-text contrasts (left) and sub-action-level pose-text contrast.

3.2.3 Refinement Supervision

The cross-modality learning described above may also introduce errors and distracting information. In P2V, the errors in pose estimation caused by unusual body poses and self-occlusions may lead to wrong visual attention. In V2P, the rich visual information may introduce distractions in pose representation. Therefore, in each sub-block, we propose to align features before and after their respective refinement to balance the inter-modal and intra-modal contrastive learning. As shown in Figure 2, we optimize the distances between video features and refined video features (i.e., z_v and z_v^R), and pose features and refined pose features (i.e., z_p and z_p^R). We establish symmetric similarities based on cosine distances, by employing the normalized temperature-scaled cross-entropy (NT-Xent) loss:

$$\mathcal{L}(z, z^R) = -\sum_{i=1}^{B} \log \frac{\exp(sim(z_i, z_i^R)/\tau)}{\sum_{j=1}^{B} \exp(sim(z_i, z_j^R)/\tau)}, \qquad (3)$$

where B is batch size, $sim(a, b) = a \cdot b^{\intercal} / (\|a\| \|b\|)$ is similarity score, and τ is a learnable temperature parameter. Finally, we combine the refinement losses from both video and pose modalities to form the contrastive loss for the UCMR Block, $\mathcal{L}_{\mathcal{R}} = \mathcal{L}(z_v, z_v^R) + \mathcal{L}(z_p, z_p^R)$.

3.3. SGMC Module

In this paper, besides exploiting an additional pose modality, we also aim to exploit the detailed text content to supervise learning among VPL modalities and alignment at both global action and fine-grained sub-action levels. Directly applying existing contrastive learning [32, 38, 50] to all pairs of global and local features from all modalities would result in a combinatorial computation explosion. We propose a novel effective strategy named *Semantic*

Guided Multi-level Contrastive (SGMC) module to achieve our goal. The SGMC module consists of two parts, of which the first is *Multi-level Contrastive Associations* among VL and PL similarities on both action and sub-action levels, and the second is a *Semantic Content Guided Loss* to conduct elastic contrast learning based on the discrepancy among text content of different fine-grained human actions.

3.3.1 Multi-level Contrastive Association

On the detailed text description of fine-grained action, the encoded text representation z_t consists of action-level sentence representation $S \in \mathbb{R}^d$ and sub-action-level word representation $W \in \mathbb{R}^{N_t \times d}$, where N_t is the number of words. The pose-adapted vision representation is learned on the whole video with a low frame rate for computational efficiency. Given the video embedding z_v^R, we obtain clip-level video representation $V \in \mathbb{R}^d$. The contrastive association between the video V and sentence S is computed at the action level, denoted as $S_{VS} = sim(V, S)$.

Pose contains concise latent semantic representation from a high frame rate, and the text contains concept-level semantic representation of fine-grained actions. The contrastive associations between these two modalities are computed on both action and sub-action levels. Given the pose embedding z_p^R, we obtain clip-level pose representation $P \in \mathbb{R}^d$, and frame-level pose representation $F \in \mathbb{R}^{T \times d}$. The similarity between pose representation P and sentence S on the action level is computed as $S_{PS} = sim(P, S)$. On the sub-action level, we compute the contrastive association matrix between frame-level pose representations F and word-level text representations W, which is denoted as $\hat{S}_{FW} = FW^{\intercal}$. Subsequently, we carry out attention operations on the matrix twice [38], to obtain the sub-action level pose-text contrastive association, S_{FW}.The overall combined similarity ($\mathbf{S}$) among VPL modalities is defined as:

$$\mathbf{S}(x_i, y_j) = \mathbf{S}((vp)_i, t_j) = (S_{VS} + S_{PS} + S_{FW})/3. \quad (4)$$

3.3.2 Semantic Content Guided Loss

Label text in fine-grained action datasets provides a detailed description of sub-actions, where two action classes may have the same sub-actions. Intuitively, a label text with a larger discrepancy from the ground-truth label text will yield a larger discrepancy from the corresponding video. Hence, simply treating a label text or action category name as either a positive or negative sample may not be accurate in learning fine-grained action context. We propose a novel *Semantic Content Guided Loss* which employs **strength coefficients** $\{s_i\}_{i=1}^B$ to adjust the pushing strength on negative samples according to the discrepancy magnitude among label texts. The strength coefficient s_i is produced by computing the cosine similarity between the sampled text (denoted by t_s) and the ground truth text (denoted by t_g): $s_i = norm(1 - sim(t_s, t_g))$, where sim denotes similarity value and $norm$ denotes normalization operation such that $\sum_{i=1}^B s_i = 1$. Then, the video&pose-to-text $(vp2t)$ and text-to-video&pose $(t2vp)$ similarity scores can be formulated by adjusting the pushing force of negative pairs with the strength coefficients $\{s_i\}_{i=1}^B$, as below:

$$p_i^{vp2t} = \frac{e^{S((vp)_i, t_i))/\tau}}{e^{S((vp)_i, t_i)/\tau} + (B-1)\sum_{\substack{j=1 \\ j \neq i}}^B s_i \cdot e^{S((vp)i, t_j)/\tau}},$$
$$p_i^{t2vp} = \frac{e^{S((vp)_i, t_i)\tau}}{e^{S((vp)_i, t_i)/\tau} + (B-1)\sum_{\substack{j=1 \\ j \neq i}}^B s_i \cdot e^{S(t_i, (vp)_j)/\tau}}. \quad (5)$$

Let q^{vp2t} and q^{t2vp} represent the ground-truth similarity scores. We balance the negative term with $B - 1$ to avoid it being too small compared with the vanilla contrastive loss. The presence of multiple videos belonging to one label in a batch makes the conventional view of learning as a 1-in-N classification problem using cross-entropy loss inappropriate. To address this issue, we adopt the Kullback–Leibler (KL) divergence as our semantic-guided loss for the training set $\mathcal{D}$ as:

$$\mathcal{L}_t = \frac{1}{2}\mathbb{E}_{(vp,t)\sim\mathcal{D}}[\mathrm{KL}(p^{vp2t}, q^{vp2t}) + \mathrm{KL}(p^{t2vp}, q^{t2vp})]. \quad (6)$$

3.4. Objective Function

To leverage the information in the three modalities (video, pose and text), we train the model with the objective function as $\mathcal{L}_{total} = \mathcal{L}_t + \lambda_{\mathcal{R}}\mathcal{L}_{\mathcal{R}}$, where $\lambda_{\mathcal{R}}$ is a hyperparameter. During testing, when provided with a raw video input x and a text y selected from a predefined label set $\mathcal{Y}$, we formulate the task as estimating conditional probability $P(S(x, y)|\theta)$, where θ denotes the model parameters. Subsequently, the testing procedure is akin to a matching procedure, where the classification result is determined by identifying the label with the highest probability score as:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} P(S(x, y)|\theta). \quad (7)$$

4. Experiments

We conducted extensive evaluations on four representative datasets for fine-grained human action recognition. The details of the experiments are presented below.

4.1. Implementation Details

Preparation: The localization of human body joints can be achieved by exploiting pose estimation methods [4, 15, 46]. Therefore, we leverage an off-the-shelf pose estimation model HRNet [46] to localize the human body joints in a video if the dataset does not provide pose data. Our text encoder g_t follows that of CLIP [41] which is a 12-layer, 512-wide Transformer with 8 heads. We use ViT (ViT-B/16 or ViT-L/14) pre-trained by CLIP for video encoder g_v and pose encoder g_p, which share weights on frozen layers. The video input spatial resolution is 224 $\times$ 224 in all the experiments. The hyper-parameter $\lambda_{\mathcal{R}}$ is set empirically to 0.1. The "Frames" column in all tables indicates the number of sampled frames used for video and pose inputs, respectively.

Training & Testing: We employ AdamW optimizer with a base learning rate of 8×10^{-6} in training. Models are trained with 50 epochs, and the weight decay is 5×10^{-2} on Tesla V100 32G GPU server. The learning rate is warmed up for the first 3 epochs and decayed to zero according to a cosine schedule for the remaining training epochs. In testing, we take the test video as input, and feed all the label texts to the model. We follow [59] by adopting multi-view inference with three spatial crops and one temporal clip for the best performance model. The final prediction is derived from the average similarity scores computed across all views, and identifying the highest score label using Equation (7).

Datasets: We use four representative fine-grained action recognition benchmark datasets: Diving48 [29], FineGym [45], HAA500 [9], and Toyota-Smarthome [10]. Diving48 encompasses 15.9K training videos and 2K validation videos, focusing on 48 diving actions. FineGym consists of two tasks: Gym99 with 20k training and 8.5k evaluation videos for 99 actions, and Gym288 with 23k training and 9.6k evaluation videos for 288 actions. HAA500 comprises 10k video clips for 500 distinct human-centric actions, distributed across training (8k), validation (500), and testing (1.5k) sets. Toyota-Smarthome encompasses 16.1k video clips featuring 18 subjects, 7 camera viewpoints, and 31 action classes within real-world scenarios, and it offers 2D skeletal representations via 13 body joints. We train our models using the standard splits and follow the established evaluation procedure. We apply standard classification accuracy as the performance metric. Our methods are evaluated across all mentioned datasets, with ablation studies conducted on Diving48.

4.2. Comparison with State-of-the-art Methods

We compare our proposed PeVL method with SOTA models on four fine-grained action recognition benchmarks. Tables 1, 2, and 3 show that our method consistently surpasses the SOTA methods. On the Diving48 dataset (see Table 1), our PeVL achieves 91.9% and 92.5% using backbones ViT-B/16 and ViT-L/14 respectively, outperforming AIM [59] with the same backbones by 3.0% and 1.9%, respectively, using lower video frame rate than AIM. When compared to ORViT [20] which leverages an object tracking model, our PeVL ViT-L/14 outperforms it by 4.5% with 51M less tunable parameters. This demonstrates the effectiveness of incorporating pose representation for cross-modality learning on fine-grained human actions. On the Gym99 and Gym288 datasets (see Table 2), our method outperforms all previous methods even when compared with RGBPose-Conv3D [13] which also takes video and pose inputs. This suggests that our model is more effective in learning human action

Table 1. Comparison to SOTA on Diving48.

Method	Tunable Param (M)	Frames	Top-1
TQN [61]	-	all	81.8
VideoSwin-B [36]	88	-	81.9
BEVT [52]	88	-	86.7
SIFAR-B-14 [14]	87	-	87.3
GC-TDN [18]	27.4	16	87.6
ORViT TimeSformer[20]	160	32	88.0
AIM ViT-B/16 [59]	11	32x3	88.9
AIM ViT-L/14 [59]	38	32x3	90.6
PeVL ViT-B/16	42	16+48	91.9
PeVL ViT-L/14	109	16+48	**92.5**

Table 2. Comparison to SOTA on Gym99 and Gym288.

Method	Gym99		Gym288	
	Top-1	Mean	Top-1	Mean
TSN [49]	86.0	76.4	79.9	37.6
TRNms [63]	87.8	80.2	82.0	43.3
TSM [31]	88.4	81.2	83.1	46.5
MTN [25]	91.8	88.5	-	-
TQN [62]	93.8	90.6	89.6	61.9
SlowOnly [17]	93.9	90.6	86.8	51.2
3D VE [26]	94.0	90.5	-	-
VT CE [26]	94.6	91.4	90.1	62.6
PoseConv3D [13]	93.2	-	-	-
RGBPose-Conv3D [13]	95.6	-	-	-
PeVL ViT-B/16	96.5	91.6	90.5	63.9
PeVL ViT-L/14	**97.0**	**91.8**	**91.8**	**64.0**

Table 3. Performance to SOTA on HAA500 (left) and Toyota-Smarthome (right).

Method	Top-1	Method	Mean
TSN [48]	55.3	AssembleNet++ [44]	63.6
CLIP [41]	63.3	MotionFormer [40]	65.8
EVL [33]	76.4	TimeSformer [2]	68.4
P2S + EVL [30]	79.8	Video Swin [34]	69.8
Semi-supervised [39]	80.7	MMNet [60]	70.1
DC-TBAC-CSN [56]	83.7	VPN++ [11]	71.0
		PAAT [43]	72.5
PeVL ViT-B/16	84.3	PeVL ViT-B/16	73.3
PeVL ViT-L/14	**84.7**	PeVL ViT-L/14	**73.8**

Table 4. Effectiveness of proposed components. Notations v,t, and p represent video, text, and pose modalities used.

Methods	**Modality**	**Top-1**	**Top-5**
Baseline	v, t	60.2	79.0
+ Pose	v, p, t	80.3	93.5
+ Adapters		87.3	97.8
+ UCMR Block		90.6	99.3
+ SGMC Module (**Ours**)		**91.9**	**99.6**

Table 5. P2V and V2P in UCMR Block.

Methods	**Top-1**	**Top-5**
PeVL	91.9	99.6
w/o UCMR Block	88.9	98.9
w/o P2V	91.1	99.3
w/o V2P	90.4	99.2
w/o Refinement Supervision	91.0	99.0

representation for fine-grained actions. It's worth noting that, in HAA500 and Toyota-Smarthome datasets, the scene background might provide spurious cues for the action class than that in Diving48 and FineGym datasets. Our PeVL can still achieve better performances compared with existing SOTA methods, as shown in Table 3. These results suggest that our method with additional pose and textual semantics can learn more effective action representations and robust to distracting visual cues. We also investigate the benefits of our model on a coarse-grained human action recognition dataset, K400 [23], detailed in supplementary materials.

4.3. Ablation Studies

We perform extensive ablation studies on our model and discuss the intriguing properties observed through the empirical results. Models in this section employ ViT-B/16 as the backbone, utilize 16-frame video input for the video encoder and 48-frame pose input for the pose encoder, and experiment on Diving48 unless specified otherwise.

4.3.1 Effectiveness of Proposed Main Components

In Table 4, we show the effectiveness of our proposed components by gradually adding them to the baseline model. We adopt the frozen CLIP model with a trainable projection layer as the baseline, forming video and text encoders with a 16-frame video input. Next, we add a frozen pose encoder consisting of temporal and spatial attention with a 48-frame pose input. Video and pose embeddings from encoders are concatenated. Next, we add trainable adapters mentioned in Section 3.1. Subsequently, we add UCMR Block and SGMC Module, respectively. It is worth noting that UCMR Block and SGMC Module can improve simple three-modality design from 87.3% to 91.9%.

Table 6. Different contrastive learning methods.

Similarity			**Top-1**
S_{PS}	S_{VS}	S_{FW}	
✓			91.0
	✓		90.9
		✓	90.9
	✓	✓	91.4
✓	✓		91.3
✓		✓	91.4
✓	✓	✓	91.9

4.3.2 Effectiveness of components in UCMR Block

To investigate the effectiveness of components in Unsymmetric Cross-Modality Refinement Block, we examine the following versions of our model: (1) w/o UCMR Block: we remove UCMR Block entirely; (2) with UCMR Block but we remove the three components inside separately: w/o P2V; w/o V2P; and w/o Refinement Supervision. As shown in Table 5, after removing UCMR Block, we observe a performance decrease from 91.9% to 88.9%. We further investigate the impact of cross-modality learning in UCMR Block. If we only remove P2V, we observe the performance drops by 0.8%. Removing V2P where the video tokens are used in enriching pose features, causes a decreased performance by 1.5%. This demonstrates both P2V and V2P help in capturing strong spatio-temporal interaction between pose temporal dynamics and video spatial representation of human actions. Furthermore, removing Refinement Supervision causes a performance drop by 0.9%. This shows that Refinement Supervision enhances the robustness of cross-modality learning with P2V and V2P.

4.3.3 Effectiveness of components in SGMC Module

To comprehensively evaluate the impact of various contrastive learning strategies, we conduct an ablation study comparing different configurations of Multi-level Contrastive Association, as outlined in Table 6. Notably, as the number of contrastive association functions increases, there is a consistent trend of improved accuracy. The highest Top-1 accuracy is obtained when PeVL incorporates all proposed contrastive associations, which underscores their synergistic effect on enhancing recognition performance. Therefore, we conclude that multi-level contrastive learning and multi-modal contrastive learning complement each other for boosted performance. Furthermore, we replace Semantic Content Guided Loss with vanilla contrastive loss by replacing strength coefficients with 1 in Equation 5, and we notice a performance drop from 91.9% to 91.4%. This shows Semantic Content Guided Loss can utilise varying discrepancies with different texts to adjust the pushing force

Table 7. Effectiveness of VPL modalities. All encoders are equipped with trainable adapters.

Method	Frames	GFLOPs	Tunable Param (M)	Top-1
V encoder	16	275	7	62.9
P encoder	48	108	11	82.6
V+P encoders	16+48	393	22	86.2
V+T encoders	16	364	13	65.1
P+T encoders	48	197	17	83.8
V+P+T encoders	16+48	482	28	87.3
PeVL	16+48	510	32	**91.9**

Table 8. Effectiveness of the adapter: "T" means having trainable adapters, otherwise, "F" means the encoder parameters are frozen.

Methods			Top-1	Top-5
Video	Pose	Text		
F	*F*	*F*	84.9	96.6
T	*F*	*F*	85.5	96.9
F	*T*	*F*	86.4	97.4
F	*F*	*T*	86.2	97.2
T	*T*	*F*	87.8	98.0
T	*T*	*T*	91.9	99.6

between the video and text, for optimal contrastive learning.

4.3.4 Effectiveness of Input Encoders

Effectiveness of VPL modalities: We investigate the effectiveness of multi-modalities in PeVL. The outcomes are presented in Table 7. Except for PeVL, video and pose modalities in rest methods are simply concatenated, and vanilla contrastive learning is adopted for text modality. Experiment results show that more modalities contribute to better Top-1 accuracy. Compared with Method "V+P+T", our PeVL achieves 91.9% Top-1 accuracy, an improvement of 4.6%. This demonstrates that our proposed framework helps learn more discriminative fine-grained representations. It's worth noting that, our PeVL outperforms SOTA method AIM [59] (91.9% v.s. 88.9%) with fewer GFLOPs (510 v.s. 809) using the same backbone architecture (ViT-B/16). This demonstrates that our cautious design on the integration of different frames of video and pose can achieve better performance more cost-effectively.

Effectiveness of Adapters: We demonstrate the effectiveness of trainable adapters by separately removing them from the video encoder, pose encoder and text encoder. The results are presented in Table 8. When all encoders remain frozen and only the UCMR Block is trained, the Top-1 accuracy is 84.9%. Performance gains with more adapters added into encoders. When all the encoders are adapted with trainable parameters, we obtain performance improvement by 7%. Conversely, the absence of adapters in any encoder leads to a negative impact on accuracy. This emphasizes the indispensability of adapters for the precise fine-tuning of downstream datasets with low-cost learning, a conclusion in alignment with our perceptual intuition.

Table 9. Ablation of the pose input forms

Input Form	GFLOPs	Tunable Param (M)	Top-1	Top-5
Heatmaps	311	11	80.1	96.0
Coordinates	108	11	82.6	96.7

4.3.5 Effectiveness of Body Joint Coordinates

In this section, we study the effect of using different pose input forms for our pose encoder. Specifically, we compare pose heatmap and body joint coordinates using only the pose encoder (e.g., "P encoder" in Table 7). We obtain heatmaps following [13], with a spatial resolution of 112x112, and extracted 48 input frames for both input forms. Table 9 shows that using body joint coordinates achieves 2.5% better accuracy and fewer GFLOPs than heatmaps, offering a simpler and more efficient representation for pose. The better performance of coordinates over heatmaps can be attributed to the inherent characteristics of each representation: (i) body joint representation provides a latent semantic structure of the human body and movement, facilitating the capturing of relationships between body parts. Whereas for heatmaps, the structural information has to be learnt implicitly from image patches; (ii) 2D pose coordinates provide explicit spatial information about the positions of body joints, while heatmaps may introduce noise due to "patchifying".

5. Conclusion

This paper introduces a novel perspective on action recognition by reframing it as a multimodal learning problem involving video-pose-text cross-modality learning with multi-level contrastive learning over the VL foundation model. We designed a multimodal architecture to exploit the appearance information of video, structural and dynamic information of pose, and semantic concepts of language. Furthermore, we formulated a new paradigm to directly adapt powerful large-scale pre-trained VL foundation models, substantially lowering re-training expenses. Our implementation, **PeVL**, is built upon CLIP, and exhibits superior performance on the fine-grained action recognition task. For future work, we plan to exploit LLM to provide more meaningful text prompts at both action and sub-action levels to better guide the joint learning of VPL modalities. We would also extend our PeVL model to hand action recognition tasks with hand pose (skeleton or joints) modality in future.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. VIVIT: A video vision transformer. In *ICCV*, 2021. 2

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 6

[3] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: Unifying the vision and language BERTs. *Transactions of the Association for Computational Linguistics (TACL)*, 2021. 4

[4] Zhe Cao, Gines Hidalgo Martinez, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[6] Santiago Castro and Fabian Caba Heilbron. Fitclip: Refining large-scale pretrained image-text models for zero-shot video understanding tasks. In *BMVC*, 2022. 2

[7] Soumyabrata Chaudhuri and Saumik Bhattacharya. Vilp: Knowledge exploration using vision, language, and pose embeddings for video action recognition. *arXiv preprint arXiv:2308.03908*, 2023. 2

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[9] Jihoon Chung, Cheng hsin Wuu, Hsuan ru Yang, Yu-Wing Tai, and Chi-Keung Tang. Haa500: Human-centric atomic action dataset with curated videos, 2021. 6

[10] Srijan Das, Rui Dai, Michal Koperski, Luca Minciullo, Lorenzo Garattoni, Francois Bremond, and Gianpiero Francesca. Toyota smarthome: Real-world activities of daily living. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2019. 6

[11] 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*, pages 1–1, 2021. 6

[12] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers, 2022. 4

[13] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2, 6, 8

[14] Quanfu Fan, Chun-Fu Chen, and Rameswar Panda. Can an image classifier suffice for action recognition? In *International Conference on Learning Representations*, 2022. 6

[15] Haoshu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2353–2362, 2016. 5

[16] Christoph Feichtenhofer, Axel Pinz, and Richard P Wildes. Spatiotemporal multiplier networks for video action recognition. In *CVPR*, 2017. 2

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 6

[18] Yanbin Hao, Hao Zhang, Chong-Wah Ngo, and Xiangnan He. Group contextualization for video recognition, 2022. 6

[19] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers, 2021. 4

[20] Roei Herzig, Elad Ben-Avraham, Karttikeya Mangalam, Amir Bar, Gal Chechik, Anna Rohrbach, Trevor Darrell, and Amir Globerson. Object-region video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3148–3159, 2022. 6

[21] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1, 2

[22] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022. 2

[23] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 7

[24] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 2

[25] Mei Chee Leong, Hui Li Tan, Haosong Zhang, Liyuan Li, Feng Lin, and Joo Hwee Lim. Joint learning on the hierarchy representation for fine-grained human action recognition. In *ICIP*, pages 1059–1063. IEEE, 2021. 6

[26] Mei Chee Leong, Haosong Zhang, Hui Li Tan, Liyuan Li, and Joo Hwee Lim. Combined cnn transformer encoder for enhanced fine-grained human action recognition, 2022. 6

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 2

[28] 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

[29] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 513–528, 2018. 6

[30] 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, 2023. 6

[31] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 6

[32] Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, and Horst Bischof. Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge, 2023. 4

[33] 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. 6

[34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3192–3201, 2021. 6

[35] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 2

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 6

[37] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18. Springer, 2022. 2

[38] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. *arXiv preprint arXiv:2208.02816*, 2022. 4, 5

[39] Xiaoyuan Ni, Sizhe Song, Yu-Wing Tai, and Chi-Keung Tang. Semi-supervised few-shot atomic action recognition, 2020. 6

[40] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12493–12506, 2021. 6

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 6

[42] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. *arXiv preprint arXiv:2212.03640*, 2022. 2

[43] Dominick Reilly, Aman Chadha, and Srijan Das. Seeing the pose in the pixels: Learning pose-aware representations in vision transformers, 2023. 6

[44] Michael S. Ryoo, AJ Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 6

[45] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[46] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation, 2019. 5

[47] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[48] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 6

[49] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE TPAMI*, 41 (11):2740–2755, 2018. 6

[50] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2, 4

[51] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 2

[52] Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Yu-Gang Jiang, Luowei Zhou, and Lu Yuan. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14733–14743, 2022. 6

[53] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. Spatiotemporal pyramid network for video action recognition. In *CVPR*, 2017. 2

[54] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. *Proceedings of the AAAI, Washington, DC, USA*, pages 7–8, 2023. 2

[55] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision, 2022. 4

[56] Santosh Kumar Yadav, Shreyas Bhat Kera, Raghurama Varma Gonela, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. Tbac: Transformers based attention consensus for human activity recognition. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2022. 6

[57] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 2

[58] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15671–15680, 2022. 2

[59] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition, 2023. 2, 3, 6, 8

[60] Bruce Yu, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP: 1–1, 2022. 6

[61] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4486–4496, 2021. 6

[62] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *CVPR*, pages 4486–4496, 2021. 6

[63] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *ECCV*, pages 803–818, 2018. 6