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

Supplementary Material

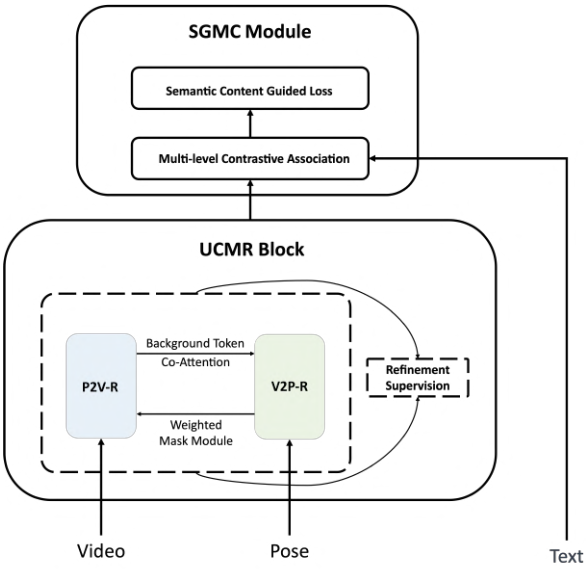

Figure 1. **Overview of PeVL**. The UCMR Block takes input video and pose tokens to derive refined video and pose tokens. These outputs are then input to the SGMC Module for multi-level contrastive learning alongside text tokens.

1. A High-level Overview

We provide a concise block diagram of PeVL in Figure 1, where each block represents a newly designed module in our paper. The details of the novel designs for each block are described in the technical part of the main paper.

2. Additional Technical Details

2.1. Encoders with adapters

Video Encoder We adopt the pre-trained image encoder in CLIP [20] with adapters [28] as the Video Encoder in PeVL. The adapter consists of two fully-connected layers and an activation layer, with a bottleneck structure. Our Video Encoder g_v adopts two adapters for spatial adaptation and output adaptation, as shown in Figure 2 (a). An advantage of this paradigm is that the network can be initialized directly from a large-scale pre-trained VL model, providing a good starting point with reasonable initial performance. Given a raw video x of a high temporal resolution T (Video (H)), we downsample x at temporal dimension to get a downsampled video of low temporal resolution T_v (Video (L)). We adopt a space-only model, to process each video frame independently. Consider the Video (L) is of size $T_v \times H \times W \times 3$, where T_v frames have a spatial resolution of $H \times W$ and three colour channels. We extract disjoint patches from each frame, resulting in video embedding v with N_v being the number of tokens. Each of these tokens is then projected to $\mathbb{R}^d$ via a linear layer. Furthermore, a [cls] token is prepended to the input sequence of each frame prior to its processing by the transformer to enable the classification of the entire video. After obtaining T_v [cls] tokens where each [cls] token stands for the representation of each frame, we added learnable position embeddings to each token to encode its position information, and subsequently pass the tokens into the video encoder for spatial modeling, denoted as *S-Attn*.

Similarly to the standard transformer, the input sequence is transformed into key, query, and value matrices denoted as $\mathbf{K_v} \in \mathbb{R}^{N_v \times d}$, $\mathbf{Q_v} \in \mathbb{R}^{N_v \times d}$, and $\mathbf{V_v} \in \mathbb{R}^{N_v \times d}$, respectively. Self-attention [23] computes the pairwise similarities between all combinations of tokens in the input sequence:

$$\boldsymbol{\alpha} = \frac{\exp(\mathbf{Q_v}\mathbf{K_v}^\intercal)}{\sum_{j=1}^{N_v} \exp(\mathbf{Q_v}\mathbf{K}_j^\intercal)} \tag{1}$$

where $\mathbf{Q_v}$, $\mathbf{K_v}$ are the d-dimensional query and key vector for tokens at the spatial position. Afterwards, we compute the average of frame features as the video representation, *i.e.* $z_v = \sum_{T_v} g_v(v)/T_v$.

Pose Encoder The pose configuration of a human in a video is typically characterized by its 2D pose, which is represented by a set of 2D coordinates (known as body joints) that provide the specific locations of various human body joints in each video frame. We pass x into an off-the-shelf 2D pose extractor to get pose representation of T frames, which is then passed through learnable embedding layers to obtain pose embedding $p \in \mathbb{R}^{T \times N_p \times d}$, where N_p is the number of body joints. Then, we pass p into the Pose Encoder g_p, which consists of a spatial modeling *S-Attn*, a temporal modeling *T-Attn*, with three learnable adapters, as shown in Figure 2 (b). We first reshape p into a size of $N_p \times T \times d$ to be fed into the *T-Attn* where it learns the relationship among the T frames. Subsequently, we reshape p into a size of $T \times N_p \times d$ to be fed into the *S-Attn* where it learns the relationship among the body joints in each frame. Though *T-Attn* and *S-Attn* take in different input dimensions, they both share the same weights and are frozen, where only newly inserted adapters are updated during training. Afterwards, we compute the

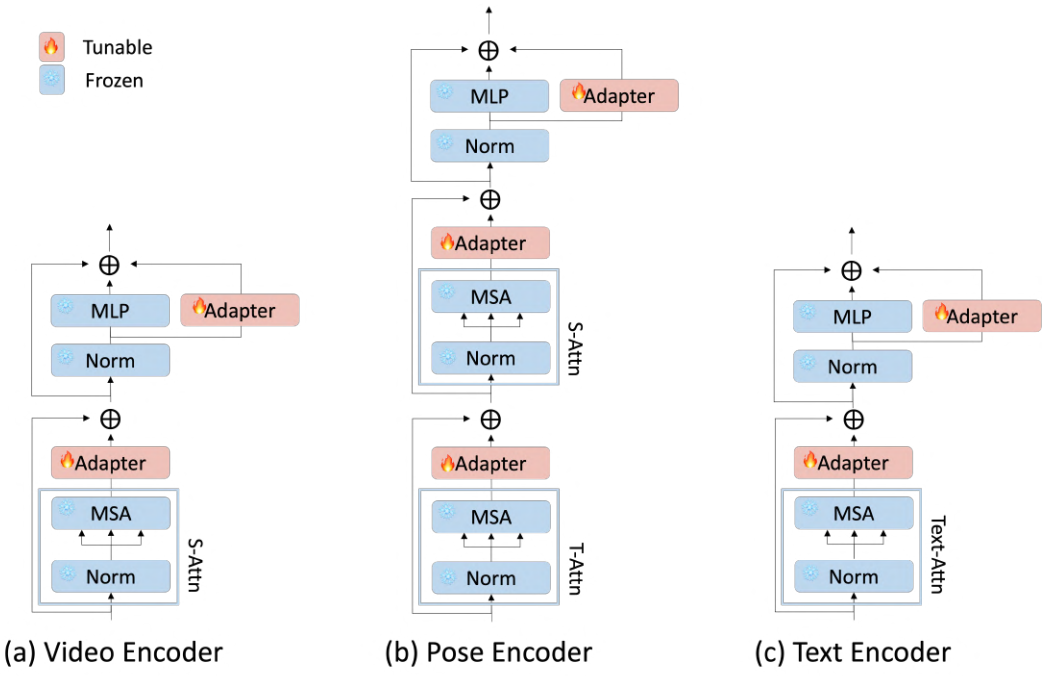

Figure 2. Encoders with trainable adapters. (a) Video encoder with adapters; (b) Pose encoder with adapters; (c) Text encoder with adapters.

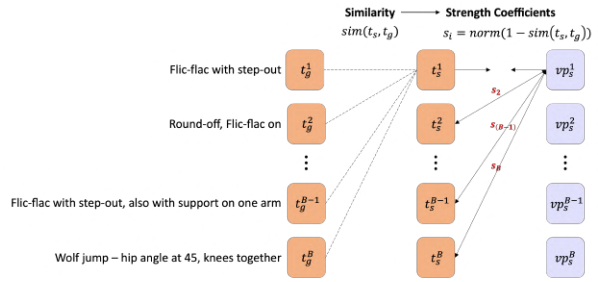

Figure 3. Strength Coefficients in Semantic Content Guided Loss.

average of frame features as the pose representation, *i.e.* $z_p = \sum_{T_p} g_p(p)/T_p$.

Text Encoder Similar to the video encoder, we adopt adapters to the CLIP text encoder. Afterwards, we pass prompted text into text encoder g_t as the text representation, *i.e.* $z_t = g_t(t')$, as shown in Figure 2 (c).

2.2. Strength Coefficients in Semantic Content Guided Loss

Figure 3 illustrates strength coefficients $\{s_i\}_{i=1}^{B}$ for a specific vp_s^j (for $j = 1$ to B, where $j \neq i$) applied in our proposed Semantic Content Guided Loss, in main paper **??**. Strength Coefficients $\{s_i\}_{i=1}^{B}$ in Semantic Content Guided Loss is to adjust the pushing strength on negative samples (t_s and vp_s) according to the discrepancy magnitude among label texts (t_s and t_g), where t_g, t_s and vp_s denote ground truth text, sampled text and sampled fused video&pose features.

2.3. Architecture of methods in ablation study of main paper

The architectures of methods mentioned in the **??**, **??** and **??** of the main paper are shown in Figure 4, Figure 5 and Figure 6, respectively. Figure 4 (a) is the same as the "VL model" in main paper **??**, and Figure 4 (e) is our proposed PeVL. In main paper **??**, the "V+P+T encoders" model is same as architecture of Figure 4 (c). In Figure 5 (c1), when "P2V" is removed, output features from the video encoder are directly fed into SGMC Module. Similarly, in Figure 5 (c2), when "V2P" is removed, output features from the pose encoder are directly fed into SGMC Module.

3. Additional Quantitative Analysis

3.1. Experiments Setting

Computational cost including tunable parameters and FLOPs for pose-extraction is not considered, as some pose data is provided by the dataset providers.

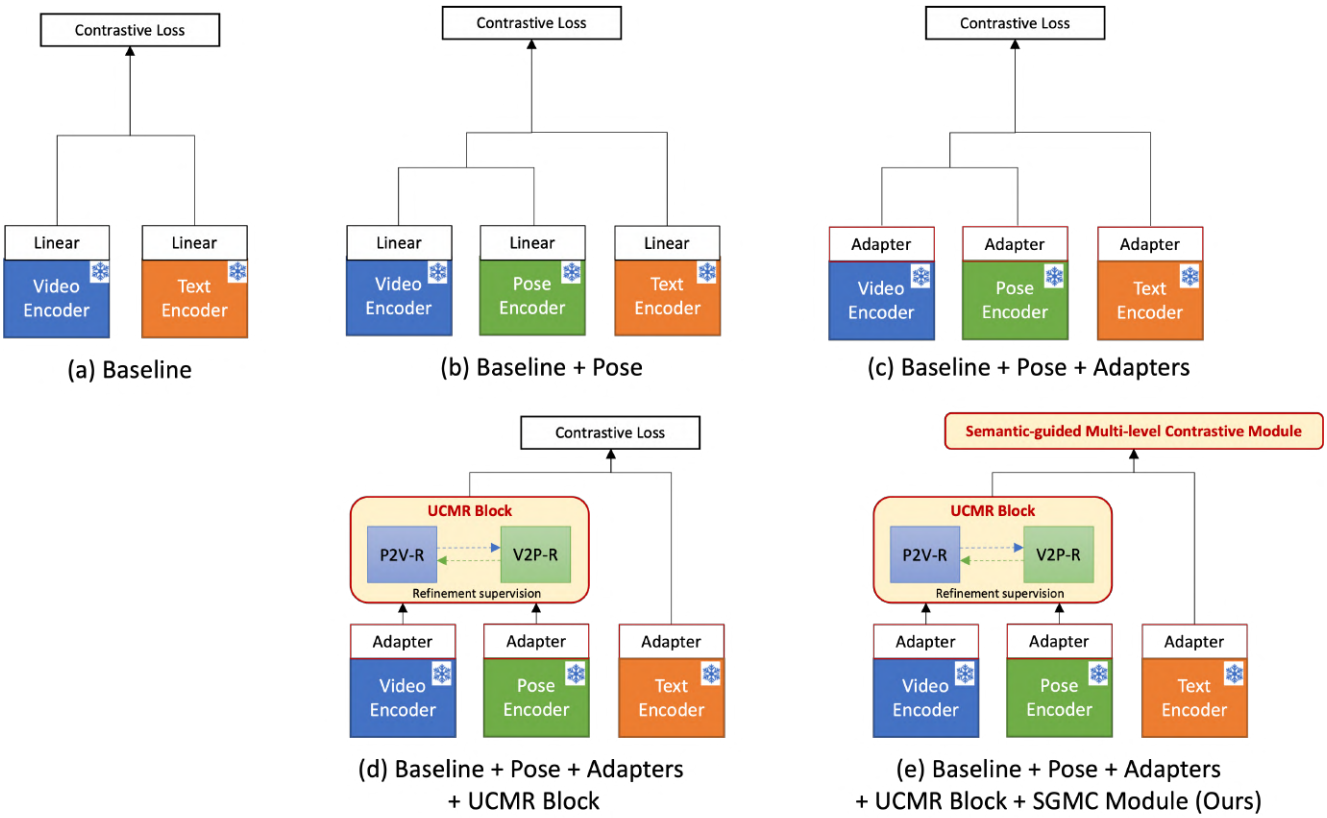

Figure 4. The architectures of methods mentioned in the main paper **??**

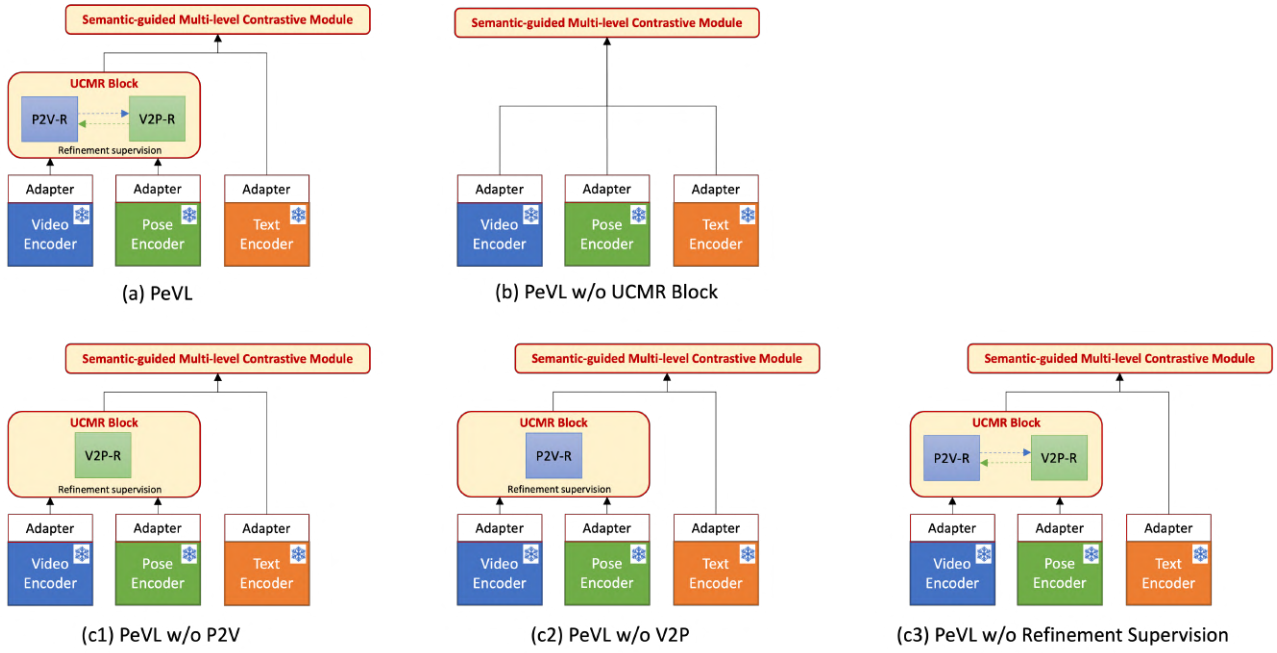

Figure 5. The architectures of methods mentioned in the main paper **??**.

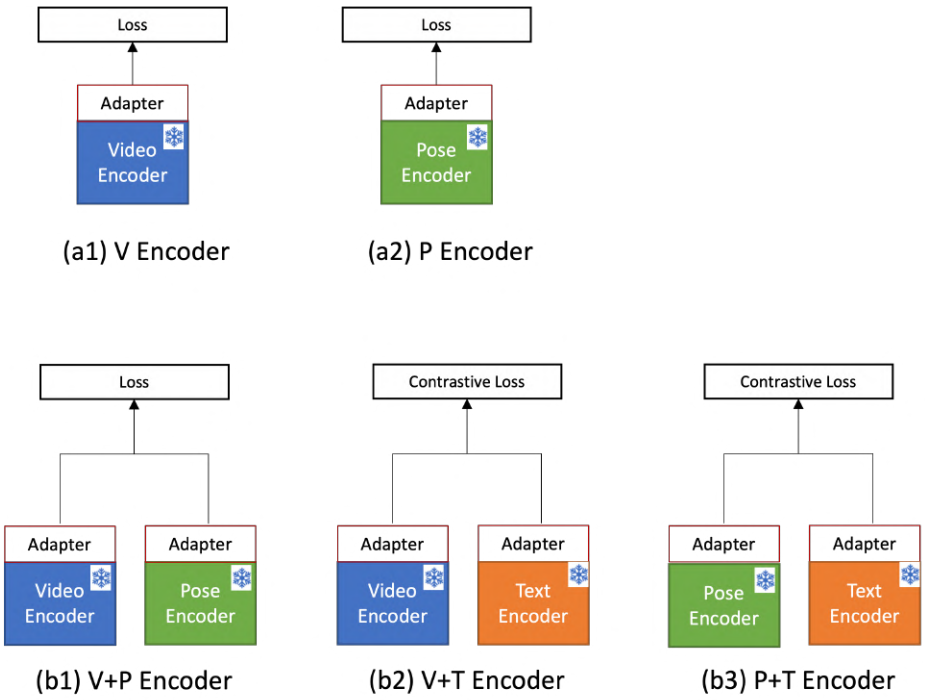

Figure 6. The architectures of methods mentioned in the main paper **??**.

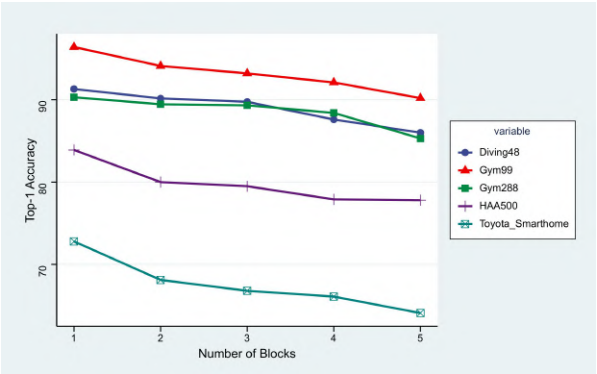

Figure 7. **Number of UCMR Blocks.** PeVL performance drops as increasing number of UCMR Blocks.

Table 1. Ablation of the pose encoder g_p

Pre-trained Model	Top-1	Top-5
Poseformer	88.0	97.9
CLIP image encoder	91.9	99.6

3.2. Additional Ablation Studies

3.2.1 Why use image encoder for pose encoder?

Given the inherent dissimilarity between images and 2D pose, our first instinct is that the model yields better results when the pre-trained data and the new data share substantial similarities. We conducted ablation studies on the pre-trained models used for the pose encoder, as shown in Table 1. Both pre-trained models are frozen and added with learnable adapters for pose encoding. If we replace the pre-trained model for the pose encoder in PeVL from CLIP image encoder to Poseformer [29], the Top-1 accuracy dropped to 88.0%. Here are some possible reasons: (1) Cross-modal information: The integration of an image vision transformer on pose data may indeed appear unconventional at first glance. While images and 2D skeletons may differ substantially, their underlying semantic representations can still provide valuable insights for action recognition. Our method is designed to bridge the gap between these modalities. By utilizing CLIP image encoder with learnable adapters for the pose modality, we can adapt the pre-trained image and text information encoded in the image encoder to learn meaningful spatio-temporal pose features. The results in Table 1 demonstrate the efficacy of our proposed approach, showcasing its ability to leverage such cross-modal information for improved performance. This further supports our hypothesis that the unique interplay of spatial and structural cues is indeed crucial in enriching the action recognition process. (2) Pre-training dataset scale: CLIP image encoder is pre-trained on a large-scale image-text dataset, whereas the PoseFormer pre-trained model we used is trained on the Human3.6M dataset [7], limiting the diversity to fine-grained action poses. The dissimilarity in the scale of the pre-training datasets plays a critical role in

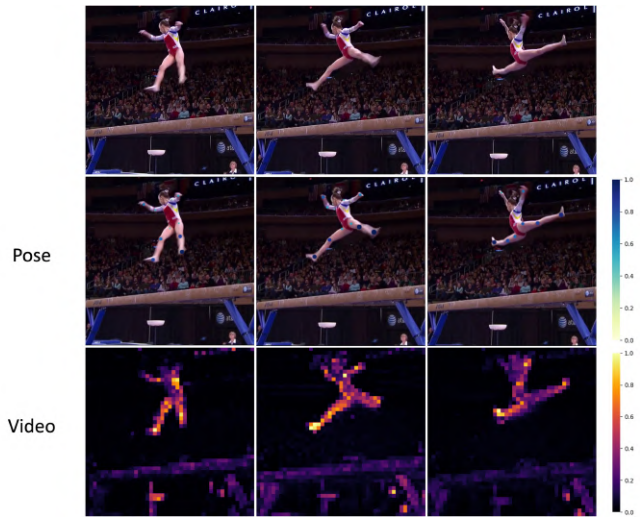

Figure 8. **Spatial Attention Visualization.** In the visualization of poses, human body joints that are attended to are represented by darker colours and larger circles compared to body joints with lower weights. For instance, in the case of the leg spreading action, there are more significant weights assigned to body joints in the lower body than in the upper body. In the visualization of videos, brighter colours indicate higher attention. Notably, we observe that greater attention is assigned to rapidly moving body parts. Moreover, spatial cues in backgrounds, such as the balance beam, also receive notable attention.

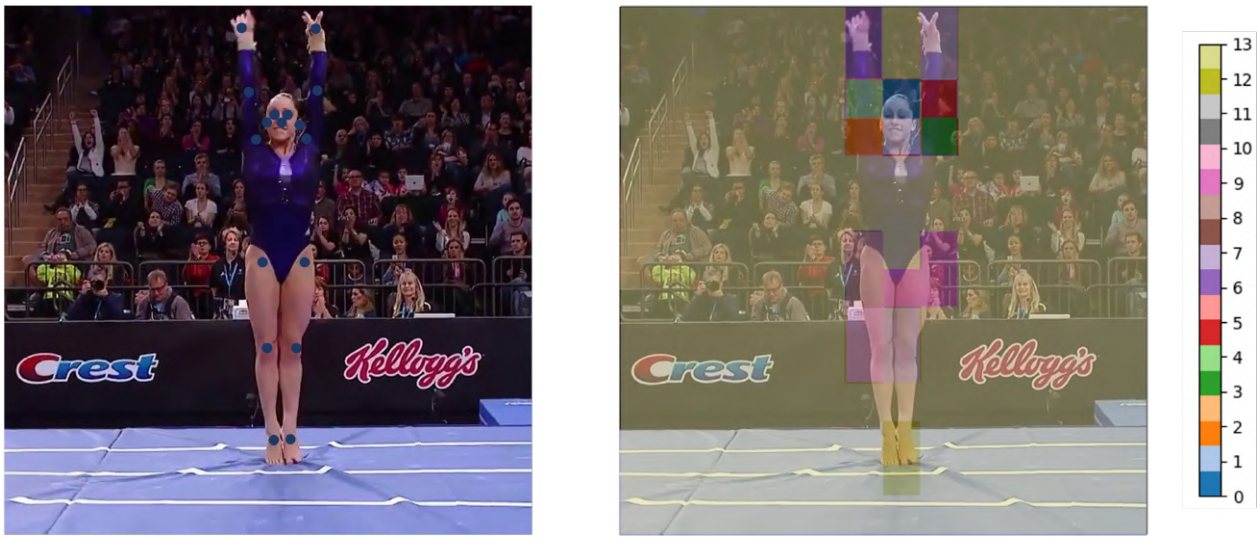

Figure 9. **Background Token Visualization.** Every image patch is attributed to either a pose token or a background token (named "grouping token" here) based on the token that exhibits the highest attention score with that particular image patch. The left image shows the location of the 17 body joints in blue; while the right image shows the grouped image patches. In this example, there are 14 grouping tokens with the highest attention scores, where the colour 13 represents image patches affiliated with the "Background Token".

determining the models' performance.

3.2.2 Temporal Attention in Pose Encoder

Having shown the effectiveness of PeVL for adapting image-text trained CLIP for fine-grained human action recognition, we explore how this approach encodes action-specific information in video that enables the bridging of the modality gap. We intend to use a small number of video frames and a relatively large number of pose frames to bridge the domain gap from the image foundation model to the video domain, for the sake of not using large GFLOPs as those pure video models. We conduct our experiments by ablating on the *T-Attn* in the pose encoder used to learn information across frames, as shown in Table 3. When we remove temporal attention layers from the pose encoder,

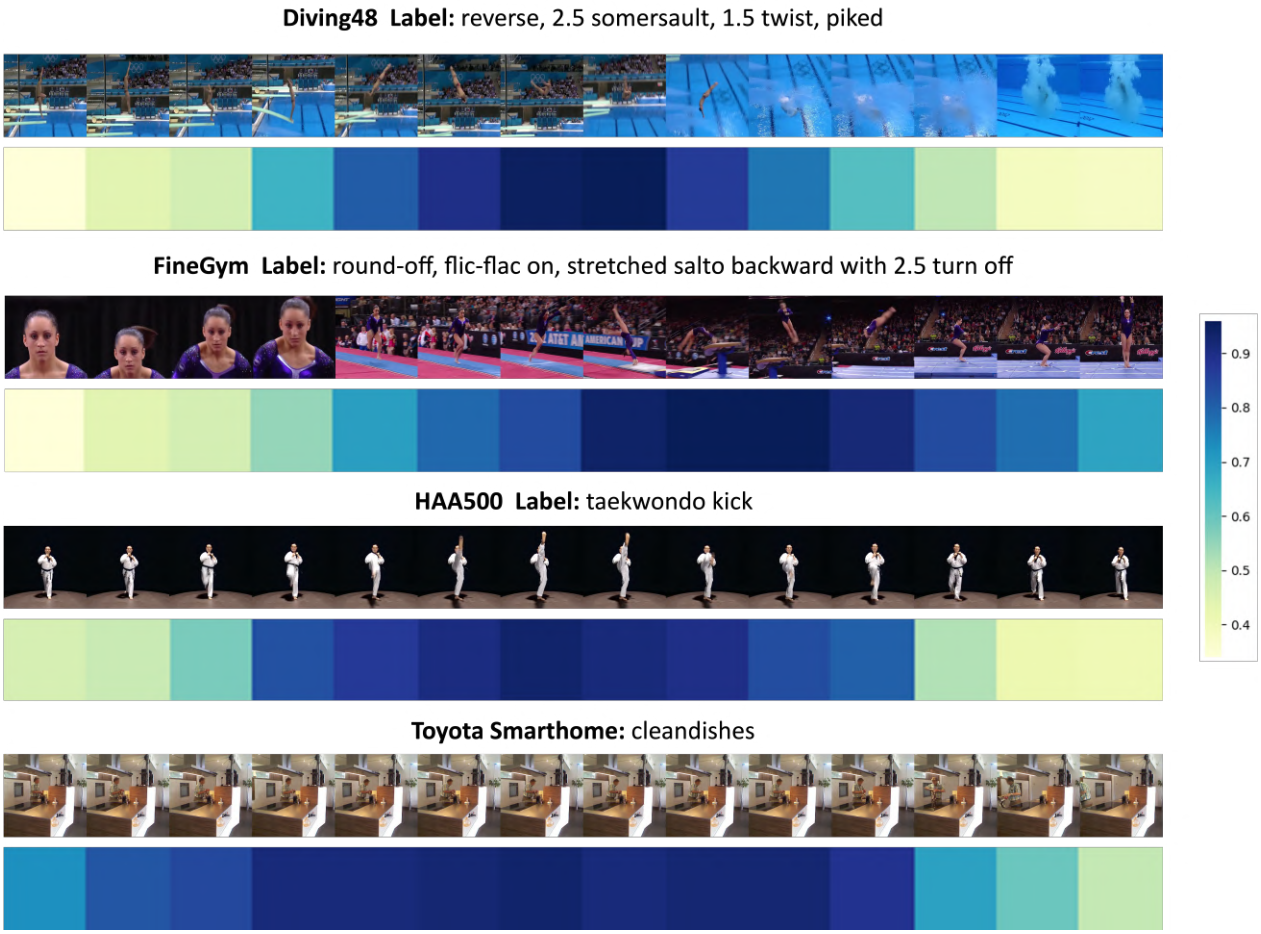

Figure 10. **Temporal Attention Visualization.** The attention patterns predicted by PeVL, displayed as coloured maps, depict temporal focus for four video clips taken from different datasets. In these maps, darker shades indicate the precise temporal positions of highlighted actions.

where both the pose and video tokens are fed directly to *S-Attn* layers, we observe a substantial performance drop of Top-1 accuracy from 91.9% to 70.0%, affirming the critical role of temporal attention in PeVL.

3.2.3 Textual Prompt

The results of the textual prompt are presented in Table 4. Notably, using textual prompts "a video of action" improves Top-1 accuracy from 91.4% to 91.7% when the sole use of label text. Incorporating coarse action type (e.g. "gymnastic" for the FineGym dataset) further improves 0.2% to 91.9%, thereby affirming the efficacy of an intelligible textual prompt with prior coarse-grained knowledge in improving performance.

3.2.4 Number of UCMR Blocks

We investigate a suitable number of UCMR Blocks employed in PeVL. Our results in Figure 7 indicate that one UCMR Block is satisfactory across all benchmark datasets, while the model's effectiveness tends to diminish with more blocks. This phenomenon can be attributed to the incorporation of additional UCMR Blocks resulting in the loss of important contextual information from non-pose tokens which also play a vital role, especially in HAA500 and Toyota-Smarthome datasets.

3.3. Performance gain on coarse-grained human action dataset

As mentioned in the main paper **??**, it is interesting to investigate the benefits of our novel model on general coarse-grained human action recognition, especially the comparison with the baseline model. Table 2 presents the comparisons with SOTA video models on the K400 [9] dataset.

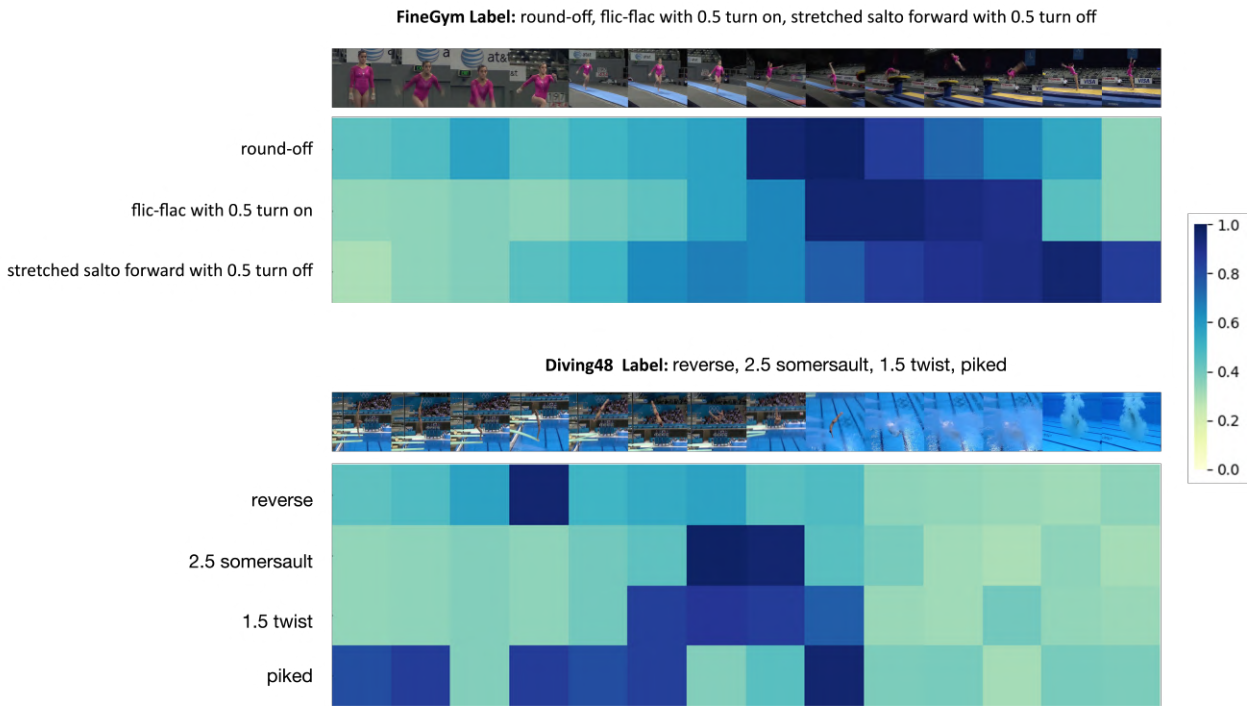

Figure 11. **Semantic Concept Attention Visualization.** For every video, we show the attention scores over 14 frames with phrases in the label text (sentence), where attention for phrases is averaged over wordings.

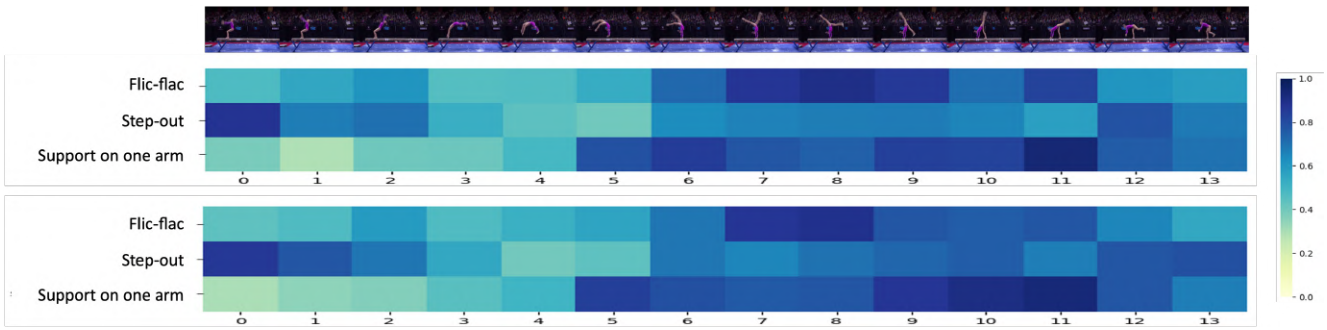

Figure 12. **Comparison Between Vanilla VL Model and PeVL on Temporal Attention.** The first row is video frames of class "Flic-flac with step-out, also with support on one arm". The second and third rows are attention visualizations of the vanilla VL model and our PeVL, respectively.

K400 contains approximately 240K training videos and 20K validation videos across 400 human action classes. According to the table, transformer-based methods achieve better results with strong vision transformers. We observe that our PeVL consumes much fewer GFLOPs and tunable parameters than most of the previous methods. PeVL ViT-L/14 achieves 89.5% top-1 accuracy using 32 pose frames and 32 video frames, which is comparable to SOTA performance, where InternVideo [26] is pre-trained on K400. When compared with the baseline AIM [28], our method outperforms by 1.3% and 2.0% in Top-1 accuracy for backbones ViT-B/16 and ViT-L/14 respectively. We attribute the comparatively modest performance on the K400 dataset to the short videos and ego-centric videos that do not have body poses, which limit the effectiveness of pose temporal reasoning. We believe that our model's strength lies in its ability to exploit the relationships between appearance, movement, and semantic concepts, particularly in scenarios where body poses are present and finer-grained text descriptions are available. The current performance of PeVL has revealed the potential of the multimodal learning framework and the proposed new paradigm on the vision foundation model for action recognition.

Table 2. Comparison to SOTA on Kinetics400.

Method	GFLOPs	Tunable Param (M)	Frames	Top-1	Top-5
TSM R50 [14]	330	24.3	8	74.1	91.2
CorrNet-101 [24]	-	-	32	79.2	-
SlowFast R101 [6]	7020	59.9	32	79.8	93.9
X3D-XXL [5]	4320	20.3	32	80.4	94.6
MoViNet-A6 [10]	386	31.4	120	81.5	95.3
MViT-B [3]	4095	37	64	81.2	95.1
UniFormer-B [11]	3108	50	32	83.0	95.4
TimeSformer-L [2]	7140	121	64	80.7	94.7
ViViT-L/16×2 FE [1]	3980	311	32	80.6	92.7
VideoSwin-L [16]	7248	197	32	83.1	95.9
MViTv2-L [13]	42420	218	32	86.1	97.0
TokenLearner-L/10 [22]	48912	450	64	85.4	96.3
PromptCLIP A7 [8]	-	-	16	76.8	93.5
ActionCLIP [25]	16890	142	32	83.8	97.1
X-CLIP-L/14 [17]	7890	420	8	87.1	97.6
EVL ViT-L/14 [15]	8088	59	32	87.3	-
MTV-L [27]	18050	876	32	84.3	96.3
Hiera-H [21]	1159x3x5	672	16	87.8	-
DualPath [18]	1868	27	32	87.7	97.8
EVA [4]	-	-	8	89.7	-
UMT-L [12]	1434x3x4	304	16	90.6	**98.7**
TubeViT [19]	17640	-	64	90.9	-
InternVideo [26]	-	1300	16	**91.1**	-
AIM ViT-B/16 [28]	2428	11	32x3	84.7	96.7
AIM ViT-L/14 [28]	11208	38	32x3	87.5	97.7
PeVL ViT-B/16	815	43	32+32	86.0	97.1
PeVL ViT-L/14	944	112	32+32	89.5	98.1

Table 3. Ablation of the *T-Attn* in pose encoder

Model	**Top-1**	**Top-5**
w/o T-Attn	70.0	81.2
PeVL	91.9	99.6

Table 4. Ablation of the textual *prompt*.

Model	**Top-1**	**Top-5**
Label Text Only	91.4	99.3
+ Textual Prompt	91.7	99.5
+ Coarse Action Type	91.9	99.6

4. Additional Qualitative Analysis

4.1. Spatial Attention

Figure 8 visualizes the weighted body joints and refined video attention for each video frame. With the proposed UCMR Block, the model learns pose-aware visual features which capture fast-moving body joints, further improving spatial recognition in fine-grained actions. Furthermore, we demonstrate the visualisation of our pose tokens with the newly introduced "Background Token" in V2P (i.e., 17 + 1 grouping tokens), as shown in Figure 9. For each image patch, we compute the attention with each grouping token and take the group label that has the highest attention. When there are overlapped joints in a single patch, the lower score labels will not be used. The inclusion of the background token ensures the effective grounding of each image patch, thereby enhancing the learning process for both action-related and appearance-related cues.

4.2. Temporal Attention

Figure 10 shows important frames along the temporal dimension. Our approach discerns and emphasizes informative frames pertinent to fine-grained recognition, disregarding non-informative ones.

4.3. Semantic Concept Attention

In essence, PeVL learns to establish temporal correspondence between semantic concepts (words) and the relevant visual features. SGMC Module is learnt and optimized to become 'experts' which can localize the corresponding frames of related concepts in the temporal feature stream. Figure 11 illustrates this action-concept correspondence, where we average the attention over words in a phrase to get the phrase's attention. In Figure 11, the upper example is taken from FineGym, which shows the model responds to "round-off" at the beginning of the action, while to "flic-flac with 0.5 turn on" and "stretched salto forward with 0.5 turn off" towards the end of the action. The lower figure in Figure 11 shows a visualization example from Diving48, where it can be seen that the attention for take-off action "reverse" is located at the beginning of the sequence, while the ones for "somersault" and "twist" span over multiple central frames, and flight position "piked" spans more frames with similar human body poses. Figure 12 shows the comparison between the vanilla VL model and our PeVL, where PeVL showcases better attention localisation on the corresponding frames. These examples show the effectiveness and advantage of SGMC Module with text concept supervised learning on video and pose features for fine-grained human action recognition.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. VIVIT: A video vision transformer. In *ICCV*, 2021. 8

[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. 8

[3] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021. 8

[4] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale, 2022. 8

[5] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 203–213, 2020. 8

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition, 2019. 8

[7] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *TPAMI*, 2013. 4

[8] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv preprint arXiv:2112.04478*, 2021. 8

[9] 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. 6

[10] Dan Kondratyuk, Liangzhe Yuan, Yandong Li, Li Zhang, Mingxing Tan, Matthew Brown, and Boqing Gong. Movinets: Mobile video networks for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16020–16030, 2021. 8

[11] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2021. 8

[12] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models, 2023. 8

[13] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4804, 2022. 8

[14] 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. 8

[15] 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. 8

[16] 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. 8

[17] 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. 8

[18] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Dual-path adaptation from image to video transformers, 2023. 8

[19] AJ Piergiovanni, Weicheng Kuo, and Anelia Angelova. Rethinking video vits: Sparse video tubes for joint image and video learning, 2022. 8

[20] 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

[21] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu

Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles, 2023. 8

[22] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: What can 8 learned tokens do for images and videos? *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 8

[23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[24] Heng Wang, Du Tran, Lorenzo Torresani, and Matt Feiszli. Video modeling with correlation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 352–361, 2020. 8

[25] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 8

[26] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning, 2022. 7, 8

[27] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, 2022. 8

[28] Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. Aim: Adapting image models for efficient video action recognition, 2023. 1, 7, 8

[29] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers, 2021. 4