Part-aware Unified Representation of Language and Skeleton for Zero-shot Action Recognition

Supplementary Material

In this supplementary material, we present additional information to explain and support our design choices and their concrete effects in each methodology module of PURLS. Additionally, we list the remaining experiment performance that we omitted in the main paper and provide a qualitative visualization of the learning outcome of our solution.

1. Additional Information for Local Semantic Divison

PURLS aligns local visual semantics with global/body-part-based/temporal-interval-based descriptions against a given label. In the submission, we meticulously analyze the spatial and temporal local semantics when they are respectively extracted, resulting in a total of $P + Z + 1$ aligned representations. Alternatively, if we jointly consider spatial-temporal semantics, there will be $P * Z + 1$ aligned representations (*i.e*. finding temporal local movements for each spatial local area). To streamline computational costs, we opt for the approach of separately considering spatial and temporal local semantics.

2. Additional Information for Creating Description-based Text Features

The descriptions for each label from every evaluation dataset are provided in the supplement repository under the path of 'supplement/gpt3_desc.xlsx'. For body-part-based local movements, we generated descriptions for four body areas including 'Head', 'Hands', 'Torso', and 'Legs'. For temporal-interval-based local movements, we generated descriptions for three phases including 'Start', 'Middle', and 'End'. A manual inspection was applied to ensure that all descriptions properly enhance the original label semantics.

2.1. Hyperparameter selection for generating descriptions

In the main submission, we used only one question and one generated answer for each description type when preparing the text encoding inputs for CLIP. Meanwhile, we also attempted to use varying numbers of questions and answers to check their influence on the model performance. In Tab. 1, we present the wrapping prompts for each global/local part and all the alternative questions we designed for each type of description generation (see the design explanation in Sec. 3.2 of our main paper). Tab. 2 records the hyperparameter ablation when we use different numbers of questions and answers for the generation. When multiple descriptions are generated, we respectively encode each answer and average all output features to calculate the linguistic embeddings later used for semantic alignments. Yet, unlike traditional supervised learning, we observed that including multiple questions and answers does not improve ZSL classification in most cases. Therefore, we only use one question and one answer for each description in our main experiments.

2.2. Preprocessing for CLIP inputs

A pre-trained text encoder from CLIP requires proper prompting on its input to ensure the backbone outputs fit the required downstream tasks. Hence, as shown in Tab. 3, we prepared customized prompting sentences to further standardize each generated description before encoding them.

3. Additional Information for Adaptive Partitioning

3.1. Hyperparameter selection for Local Representations

In the main submission, we choose to align local movements using four body-part-based (BP) partitions and three temporal-interval-based (TI) partitions (*i.e*. $P = 4$, $Z = 3$). Tab. 4 provides the hyperparameter ablation when we try different numbers of spatial/temporal local representations. In the case of BP, using one body part means directly calculating the global features along the spatial dimension. Using two body parts refers to splitting spatial features into upper and lower body movements. Using six body parts requires further decomposing hand and leg movements into individual single-hand and single-leg movements. In the case of TI, the original sequence is averagely divided into multiple time intervals according to the specified number. The results reveal that having four body parts and three temporal intervals yields the best performance for PURLS.

3.2. Visualization for Adaptive Partitioning

Fig. 1, Fig. 2, Fig. 3, and Fig. 4 visualize the learned weights for acquiring the corresponding visual representations towards the descriptions for the classes 'apply cream on face', 'running on the spot', 'kicking other person', and 'cutting nails'. We analyze the impacts of adaptive partitioning in four scenarios. Two spatial-based scenarios include learning the actions characterized by specific local body movements (*e.g*. 'applying cream on face') or the movement from

Description Type	Questions
Body-part-based	Using the following format, describe in very short **how each body part moves / how a person's body part would act / individual body part action** for <Action>.
	Head would: {*text*}
	Hands would: {*text*}
	Torso would: {*text*}
	Legs would: {*text*}
Temporal-interval-based	Use the following format, separate in very short **how a person performs the action of / the motion of / the sub-motions when a person carries out** <Action>into three phases.
	1: The person would {*text*}.
	2: The person would {*text*}.
	3: The person would {*text*}.
Global	Using the following format, describe in very short **the motion of a person who / the motion of a person carries out / how a person does** <Action>.
	The person would {*text*}.

Table 1. The full version of the designed questions used to generate label-relative descriptions under different scales. The bold texts represent the question alternatives, while {*text*} refers to the blanks for GPT-3 to fill.

# Prompt question	# Description per question	NTU-RGBD 60 (Acc %)				NTU-RGBD 120 (Acc %)			
		55/5	48/12	40/20	30/30	110/10	96/24	80/40	60/60
1	1	**79.23**	40.99	**31.05**	23.52	**71.95**	**52.01**	**28.38**	**19.63**
2	1	75.44	**45.64**	26.38	**24.72**	61.87	45.28	25.10	15.76
3	1	76.90	43.63	26.37	24.67	63.00	43.96	25.01	15.73
1	5	78.50	41.57	27.25	22.02	67.59	44.06	26.32	16.39
1	10	77.7	41.99	27.01	22.33	66.51	44.60	26.56	16.96

Table 2. Hyperparameter ablation (%) of applying different numbers of prompt questions & generated descriptions on GPT-3. '# Prompt question' refers to the number of used questions. '# Desc per question' means the number of generated descriptions for each question.

the entire body (*e.g.* 'running on the spot'). Two temporal-based scenarios include learning the actions that can be broken down into sequentially local movements (*e.g.* 'kicking other person'), and the actions that involve repetitive movements (*e.g.* 'cutting nails'). To showcase the different learning focus in a more intuitive manner, we demonstrate the respective feature sampling weight of each body joint on every temporal dimension.

Adaptive partioning for spatially local movements: For the actions characterized by specific local body movements (Fig. 1), against body-part-based descriptions (Row 1-4 in each phase), adaptive partitioning effectively highlights the feature sampling from the body joints belonging to the 'hands'. It also assigns contextual significance to a few body-joint features from the 'torso' and 'legs' during the middle phase, as some co-movements may exist when raising arms to one's face. The local representation of hand movements emphasizes these features the most. The representations for other descriptions learn each joint feature in a more averaged manner while still giving the most importance to hand features, as they become the most valuable context features. Against the actions characterized by the entire body movement (Fig. 2), we find the module more averagely samples the body-joint-level features for each description.

Adaptive partioning for temporally local movements: For the actions that can be broken down into sequentially local movements (Fig. 3), against temporal-interval-based descriptions (Row 5-7 in each phase), adaptive partitioning emphasizes the overall features in the third phase, which contains the most representative postures of the kicking movement. The representation for the 'end' phase assigns the highest weights to these features to represent its local description as 'Strike other person.' Meanwhile, multiple hand-related features are collected with higher priority in the first two phases, while some leg-related features are also collected in the second phase. This implies that PURLS also allocates attention to sampling the hand and leg movements for representing the descriptions of 'Raise leg' and 'Extend foot'. Against the actions that repeat temporally local movements (Fig. 4), we observe a relatively balanced distribution of learning focus. In particular, the distribution tends to concentrate slightly more in the first phase. We believe this is because most relevant visual information is already available at the beginning. On the other hand, the ending phase may contain noises from the zero-padding frames.

Description Type	Prompts
Body-part-based	A cropped video of people's head motions that <Description>.
	A cropped video of people's hand motions that <Description>.
	A cropped video of people's torso motions that <Description>.
	A cropped video of people's leg motions that <Description>.
Temporal-interval-based	A trimmed video of the motion that <Description>.
Global	A video of people's motion that <Description>.

Table 3. The customized prompts used to wrap each description before it is sent to CLIP.

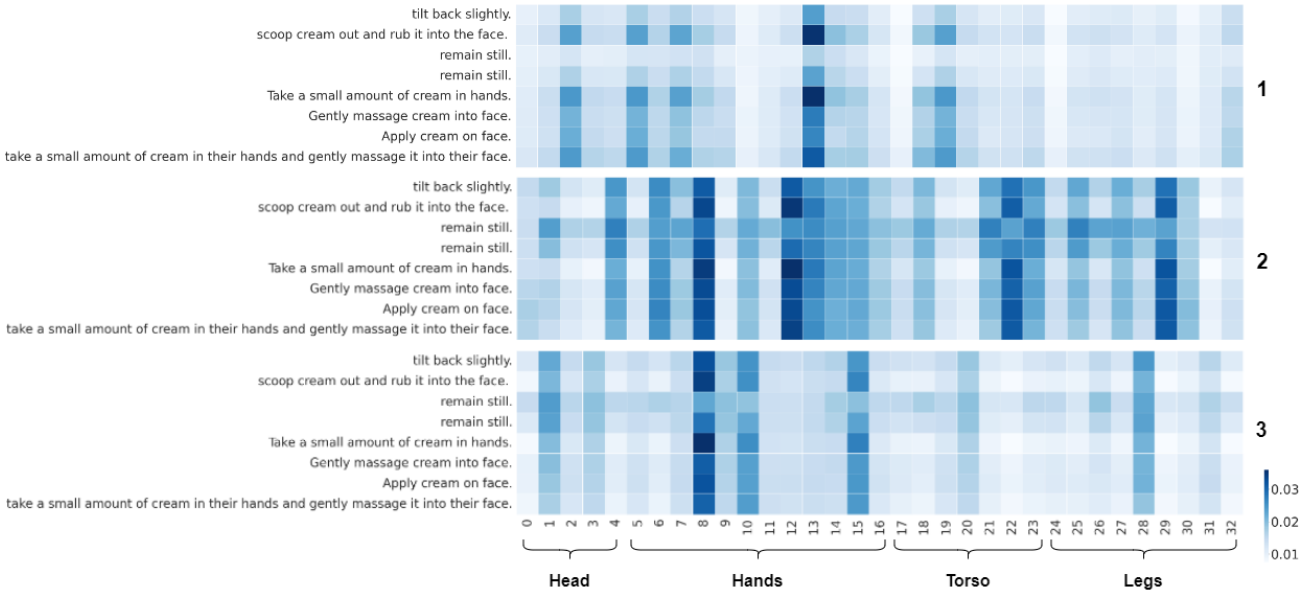

Figure 1. Visualized illustration of the learned adaptive weights used for sampling joint features to generate the visual representation of each description on 'apply cream on face'. The horizontal axis lists each body joint, and the vertical axis is the convoluted temporal dimension with a length of L'. We trunk the temporal dimension into three groups for the phases of 'start' (1), 'middle' (2), and 'end' (3). And we label the default body part that each body joint usually belongs to according to Fig. 3 in our main paper. For the demonstration in one phase, from top to bottom is the feature weight distribution against the description for body-part-based ('head', 'hands', 'torso', 'legs'), temporal-interval-base ('start', 'middle', 'end'), and global semantics.

BP	TI	NTU-RGBD 60 (Acc %)				NTU-RGBD 120 (Acc %)			
		55/5	48/12	40/20	30/30	110/10	96/24	80/40	60/60
1	3	77.70	40.69	28.84	22.46	71.26	46.13	24.43	18.57
2	3	78.31	33.15	30.81	23.03	72.77	45.90	26.26	**19.65**
4	3	**79.23**	**40.99**	**31.05**	**23.52**	**71.95**	**52.01**	**28.38**	19.63
6	3	72.18	37.65	30.10	23.35	62.67	47.81	26.39	18.78
4	1	76.32	37.62	29.06	21.91	71.73	40.92	23.49	19.13
4	3	**79.23**	**40.99**	**31.05**	**23.52**	**71.95**	**52.01**	**28.38**	19.63
4	6	74.54	38.89	28.70	23.03	71.00	48.81	26.11	17.71

Table 4. Hyperparameter ablation (%) on *NTU-RGB+D 60* and *NTU-RGB+D 120* of (1) using different numbers of body-part-based (BP) local partitioning, (2) using different numbers of temporal-interval-based (TI) local partitioning.

4. Additional Information for Experiments

4.1. Seen/unseen Split Setups

The specific lists of seen/unseen classes in each experiment split are provided as numpy files in the supplement repository under the path of 'supplement/label_splits'. Each split setting belongs to one of the following dataset folders: 'ntu60', 'ntu120', 'kinetic200', 'kinetic400', 'nw-ucla', 'utd-mhad', 'uwa3dii'. The first three datasets are the ones we focus on in the main paper. For a given split, the file that contains the corresponding unseen class list is named as 'ru + the number of unseen classes'. Similarly, the file recording the corresponding seen classes is named as 'rs + the number of seen classes'. For example, for the split with 55 seen classes and 5 unseen classes on the *NTU-RGB+D 60* dataset, the corresponding seen and unseen class lists are recorded in the files of 'supplement/label_splits/ntu60/rs55.npy' and 'supplement/label_splits/ntu60/ru5.npy'.

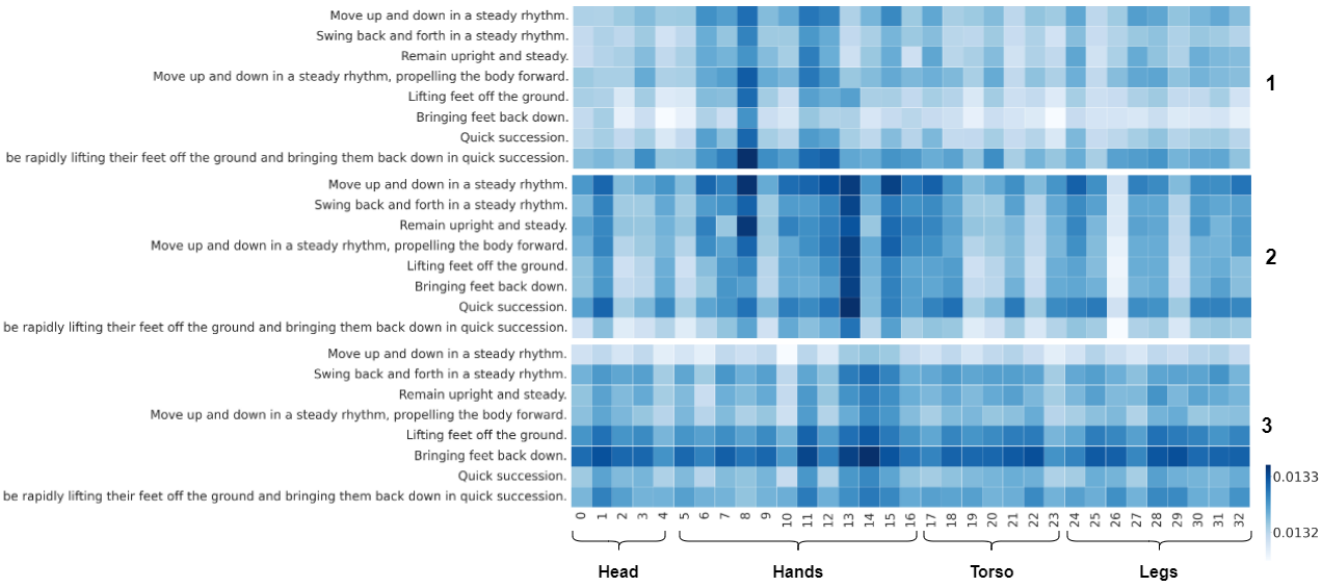

Figure 2. Visualized illustration of the learned adaptive weights for the class 'running on the spot'.

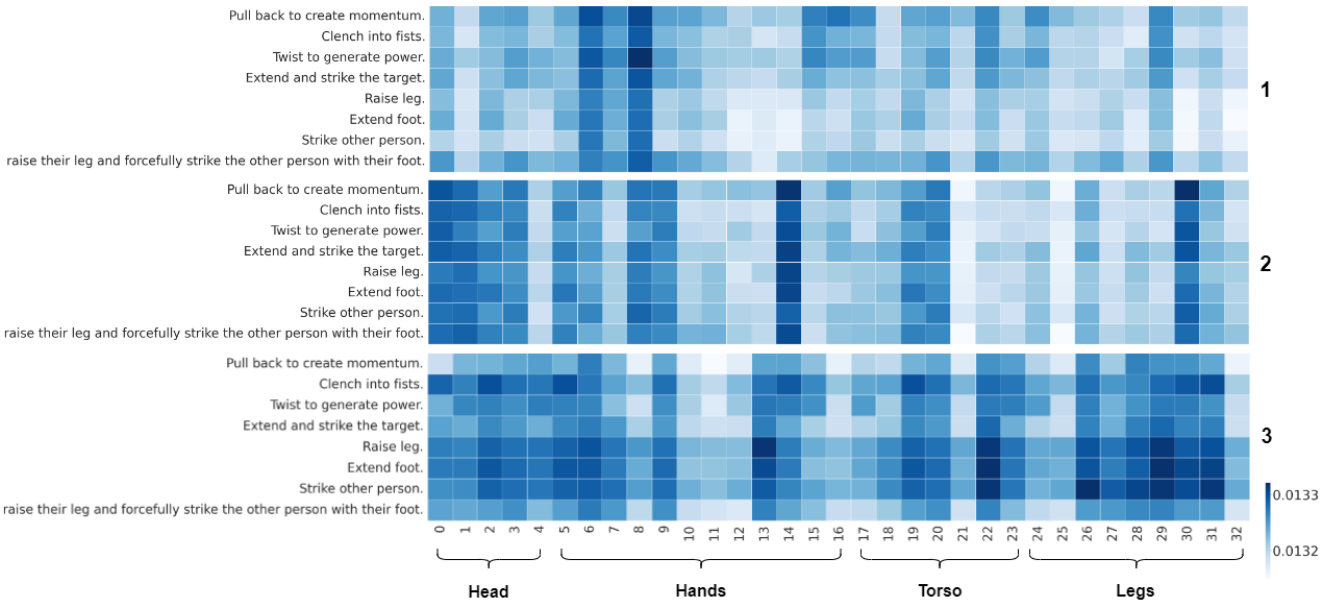

Figure 3. Visualized illustration of the learned adaptive weights for the class 'kicking other person'.

4.2. Extra Visualization for Experiment Results

Fig. 5, Fig. 6 and Fig. 7 visualize the accuracy gaps and changing curves of every baseline and PURLS when predicting different numbers of unseen classes on *NTU-RGB+D 60*, *NTU-RGB+D 120*, *Kinetic-skeletons 200*. Our method reaches state-of-the-art prediction accuracies in every experimental setting. The performance conclusion on each dataset is consistent with the analysis in Sec. 4.3 of our main paper.

4.3. Full Ablation Results

Tab. 5 and Tab. 6 add extra results for the experiment setups we used for ablation study (see Sec. 4.4 of our main paper) on the *Kinetic-skeletons 200* dataset.

4.4. Other Experiment Results

Kinetic-skeleton 400: We provide Tab. 7 to record the performance results of ReViSE, DeViSE, Global, and PURLS on *Kinetics-skeleton 400* (See the dataset description in

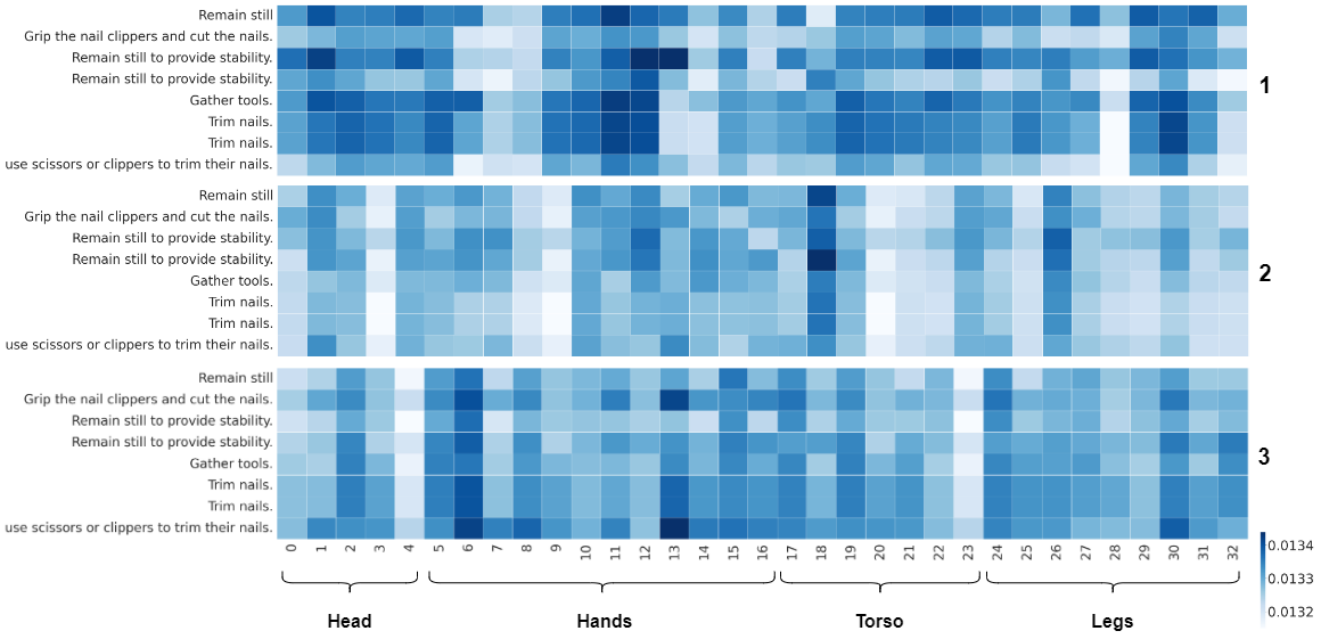

Figure 4. Visualized illustration of the learned adaptive weights for the class 'cutting nails'.

Partitioning Strategy	NTU-RGBD 60 (Acc %)				NTU-RGBD 120 (Acc %)				Kinetic 200 (Acc %)			
	55/5	48/12	40/20	30/30	110/10	96/24	80/40	60/60	180/20	160/40	140/60	120/80
Global (Original)	64.69	35.46	27.15	16.29	66.96	44.27	21.31	14.12	25.96	15.85	10.23	7.77
Global (GPT-3)	78.50	33.47	29.21	22.27	64.89	47.15	25.16	17.46	24.44	14.08	8.31	7.06
Static Partitioning	76.46	33.03	29.57	22.00	67.62	46.83	26.98	18.03	24.04	15.60	8.14	7.74
Adaptive Partitioning	**79.23**	**36.77**	**31.05**	**23.52**	**71.95**	**52.01**	**28.38**	**19.63**	**32.22**	**22.56**	**12.01**	**11.75**

Table 5. Full ablation study (%) on *NTU-RGB+D 60*, *NTU-RGB+D 120*, and *Kinetics-skeleton 200* for using different alignment learning with partitioning strategies, including direct global feature alignment to label or global description semantics, and PURLS with static/adaptive partitioning.

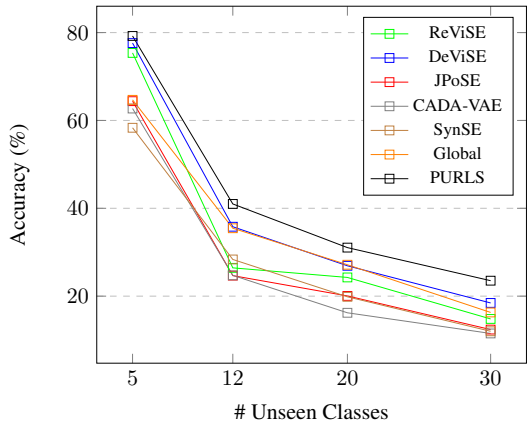

Figure 5. Visualized accuracy variation on *NTU-RGB+D 60*.

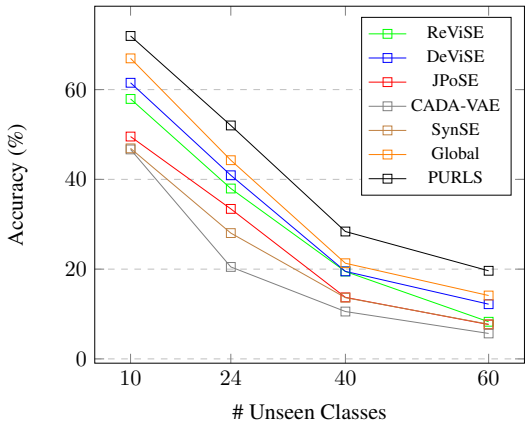

Figure 6. Visualized accuracy variation on *NTU-RGB+D 120*.

Sec. 4.1 of our main paper). Fig. 8 visualizes the accuracy gaps and changing curves of every baseline and PURLS. We created four splits of 360/40, 320/80, 300/100, and 280/120 for evaluation. These setups are replaced with

α_i	BP	TI	NTU-RGBD 60 (Acc %)				NTU-RGBD 120 (Acc %)				Kinetic 200 (Acc %)			
			55/5	48/12	40/20	30/30	110/10	96/24	80/40	60/60	180/20	160/40	140/60	120/80
-			78.50	33.47	29.21	22.27	64.89	47.15	25.16	17.46	24.44	14.08	8.31	7.06
Average	✓		76.68	37.80	30.92	22.20	68.11	30.93	24.36	18.67	22.32	7.12	3.63	5.60
Learnable	✓		76.32	37.62	29.06	21.91	71.73	40.92	23.49	19.13	22.73	14.79	8.24	7.69
Average		✓	78.65	38.80	28.14	22.69	55.73	50.67	27.50	17.50	21.81	20.04	8.61	6.81
Learnable		✓	77.70	40.69	28.84	22.46	71.26	46.13	24.43	18.57	24.85	18.73	7.97	7.95
Average	✓	✓	79.02	39.92	31.00	23.47	**73.55**	51.38	27.67	18.66	26.87	21.10	10.03	11.55
Learnable	✓	✓	**79.23**	**40.99**	**31.05**	**23.52**	71.95	**52.01**	**28.38**	**19.63**	**32.22**	**22.56**	**12.01**	**11.75**

Table 6. Full ablation study (%) on *NTU-RGB+D 60*, *NTU-RGB+D 120* and Kinetics-skeleton 200 for (1) using different α_i ($i \in [0, P + Z + 1)$) to sum for L_{train}, (2) adding body-part-based (BP) alignment learning, (3) adding temporal-interval-based (TI) alignment learning.

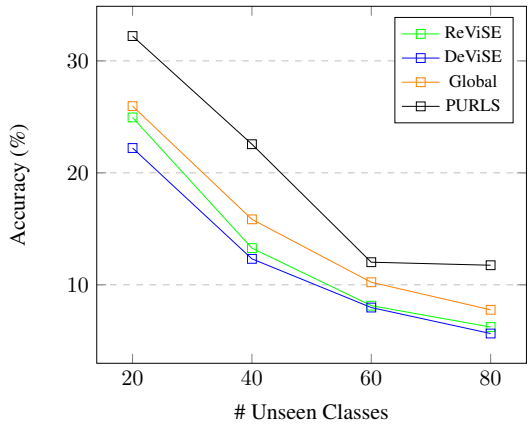

Figure 7. Visualized accuracy variation on *Kinetics-skeleton 200*.

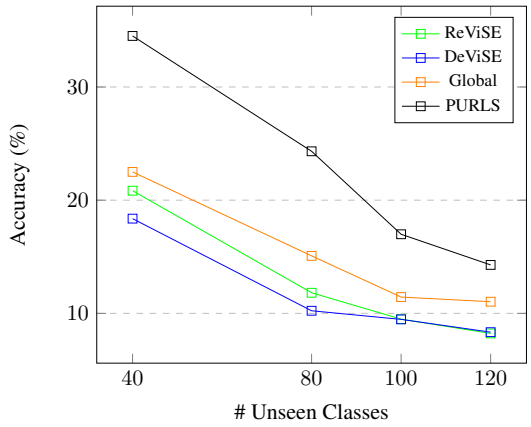

Figure 8. Visualized accuracy variation on *Kinetics-skeleton 400*.

a similar but simpler testing protocol on *Kinetics-skeleton 200*.

Small-scale datasets: Tab. 8 records the performance of ReViSE, DeViSE, Global, and PURLS on multiple small-scale datasets, including *NW-UCLA*, *UTD-MHAD*, and *UWA3D II* (respectively containing 10, 27 and 10 common daily action classes). As shown in the table, PURLS also outperforms other baselines in these testing environments where a minimal semantic overlap between seen and unseen classes makes zero-shot recognition more challenging.

Model	Kinetic 400 (Acc %)			
	360/40	320/80	300/100	280/120
ReViSE	20.84	11.82	9.49	8.23
DeViSE	18.37	10.23	9.47	8.34
Global	22.50	15.08	11.44	11.03
PURLS	**34.51**	**24.32**	**16.99**	**14.28**

Table 7. Zero-shot action recognition results (%) on *Kinetics-skeleton 400* for PURLS & all available baselines introduced in Sec. 4.2 of our main paper.

Model	NW-UCLA		UTD-MHAD			UWA3D II		
	8/2	5/5	22/5	18/9	14/13	24/6	20/10	15/15
ReViSE	69.12	44.99	29.37	19.86	12.26	30.33	10.91	10.43
DeViSE	72.81	36.02	12.50	13.59	13.94	34.16	12.51	11.41
Global	73.49	46.83	23.00	18.47	14.38	32.18	15.63	11.02
PURLS	**75.84**	**49.47**	**57.50**	**31.71**	**19.23**	**35.65**	**18.91**	**13.98**

Table 8. Zero-shot action recognition results (%) on NW-UCLA, UTD-MHAD and UWA3D II for PURLS & all available baselines.