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

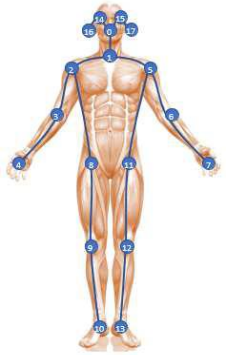
Anqi Zhu [1], Qiuhong Ke [2], Mingming Gong [1], James Bailey [1]

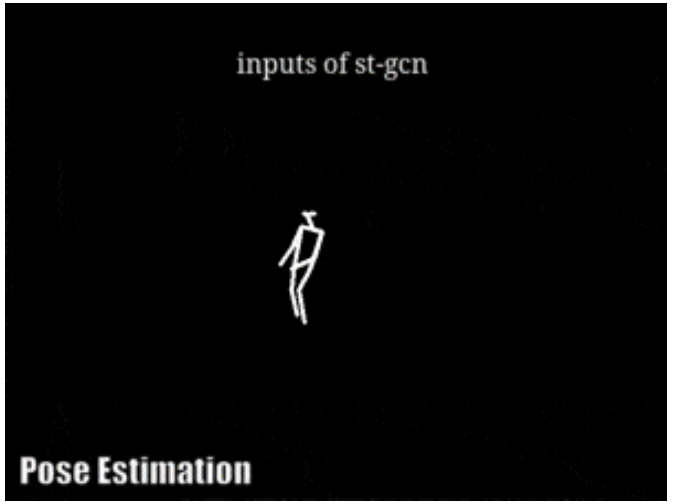
[1] University of Melbourne

[2] Monash University

Skeleton-based Human Action Recognition

Video source:
https://github.com/open-mmlab/mmskeleton

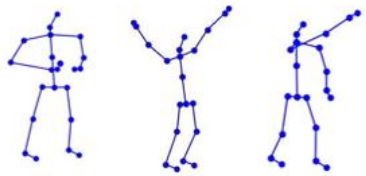




Zero-shot Learning for Skeleton-based Recognition

Training: *"Seen" Classes*

Data Inputs:



References:

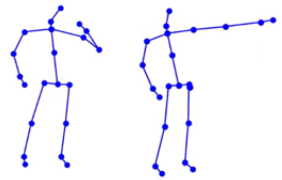
Labels + Auxiliary Info (e.g. attributes)

Model

Testing: *"Unseen" Classes*

Test Inputs:

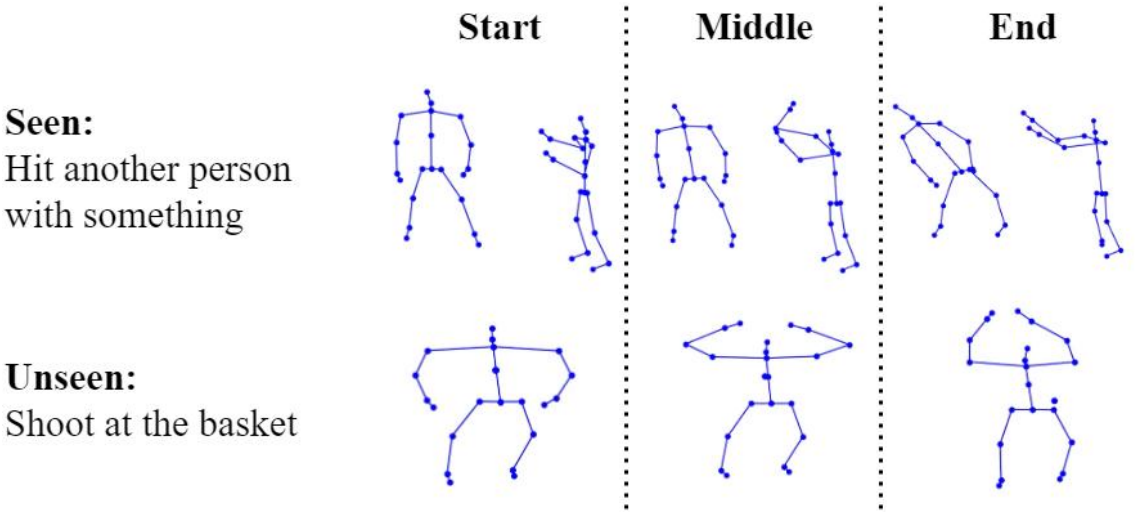
(Opt: + Auxiliary info)



Predictions:

Labels (Hush, Point fingers…)

Intuitive insights in Action-Specialized Learning Environments…



Spatial Decomposition:

Class	Head	Hands	Torso	Legs
Hit another person	Turn to the person.	Grip the object and thrust forward.	Twist and turn.	Stomp the ground.
Shoot at the basket	Turn and look up.	Grip the ball and release it.	Twist and expand.	Bend slightly and propel.

Temporal Decomposition:

Class	Start	Middle	End
Hit another person	Raise arm.	Swing arm.	Strike another person.
Shoot at the basket	Raise arm.	Throw ball.	Aim at basket.

Label Description:

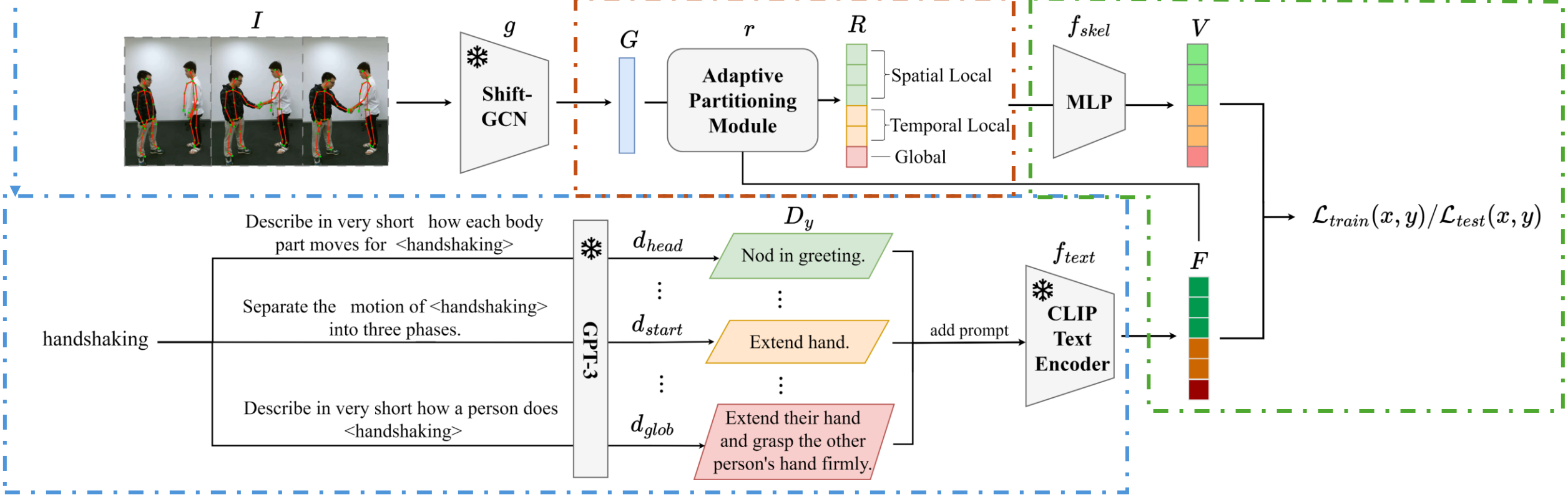
Class	Description
Hit another person	Swing their arm and strike the other person with the object.
Shoot at the basket	Raise their arm and throw the ball towards the basket.

*Same/Similar Motion Semantics

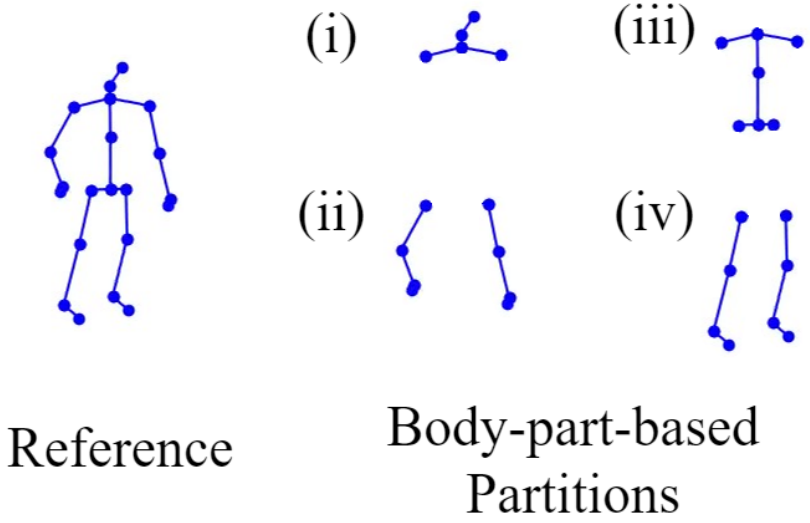
Our Contributions:

- Generate text descriptions and features for action labels & their divisible local movements.
- Sample the corresponding visual representation for each description.
- Realize global/local visual knowledge alignment & transfer.

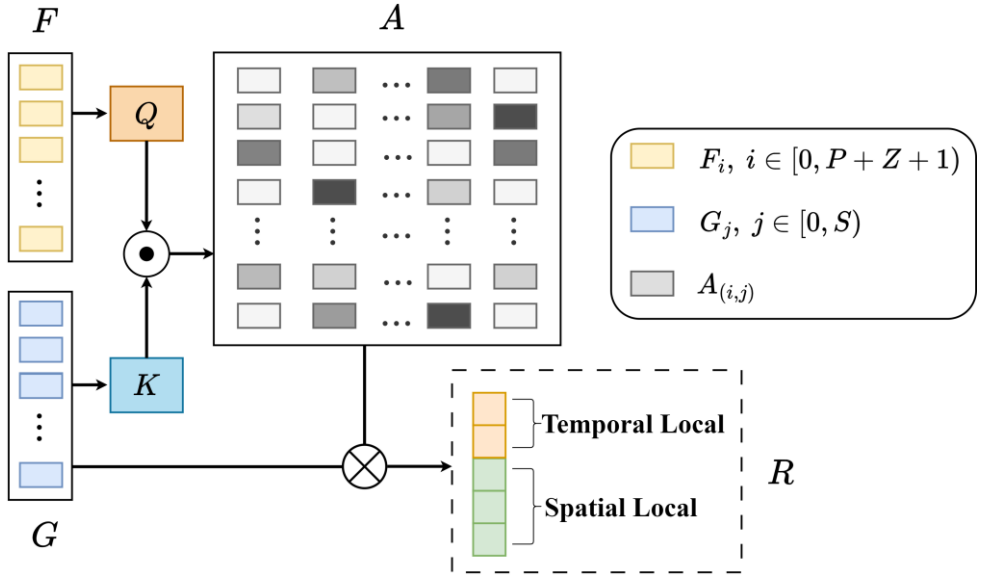
Architecture of PURLS



Design Choices for Generating Visual Representations



(a) Static Partitioning



(b) Adaptive Partitioning

Result Analysis

Model	NTU-RGBD 60 (Acc %)				NTU-RGBD 120 (Acc %)				Kinetics-skeleton 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
ReViSE [36]	75.37	26.44	24.26	14.81	57.92	37.96	19.47	8.27	24.95	13.28	8.14	6.23
DeViSE [14]	77.61	35.80	26.91	18.45	61.52	40.91	19.50	12.19	22.22	12.32	7.97	5.65
JPoSE [40]	64.82	28.75	20.05	12.39	51.93	32.44	13.71	7.65	-	-	-	-
CADA-VAE [29]	76.84	28.96	16.21	11.51	59.53	35.77	10.55	5.67	-	-	-	-
SynSE [15]	75.81	33.30	19.85	12.00	62.69	38.70	13.64	7.73	-	-	-	-
SMIE [48]	77.98	40.18	-	-	65.74	45.30	-	-	-	-	-	-
Global	64.69	35.46	27.15	16.29	66.96	44.27	21.31	14.12	25.96	15.85	10.23	7.77
PURLS	**79.23**	**40.99**	**31.05**	**23.52**	**71.95**	**52.01**	**28.38**	**19.63**	**32.22**	**22.56**	**12.01**	**11.75**

(a) Model Performance

Ablation Study

Encoder	Descriptor	Model	NTU-RGBD 60 (Acc %)			
			55/5	48/12	40/20	30/30
AA [33]	GPT3	Global	62.79	28.09	25.66	13.86
AA [33]	GPT3	PURLS	76.75	32.39	31.00	21.86
CTR [9]	GPT3	Global	65.16	34.56	26.12	15.92
CTR [9]	GPT3	PURLS	79.97	39.42	32.26	24.59
DG [32]	GPT3	Global	64.28	34.04	27.63	16.71
DG [32]	GPT3	PURLS	80.41	41.06	33.77	25.12
PoseC3D [13]	GPT3	Global	63.45	35.71	27.88	20.66
PoseC3D [13]	GPT3	PURLS	81.14	41.60	**34.47**	**28.11**
Shift	GPT3	Global	64.69	35.46	27.15	16.29
Shift	GPT3	PURLS	79.23	40.99	31.05	23.52
Shift	GPT3.5	Global	66.49	38.01	26.31	17.35
Shift	GPT3.5	PURLS	79.17	40.98	30.07	19.95
Shift	GPT4	Global	64.71	40.76	25.68	20.58
Shift	GPT4	PURLS	**81.53**	**41.90**	27.28	21.45

(b) Universality

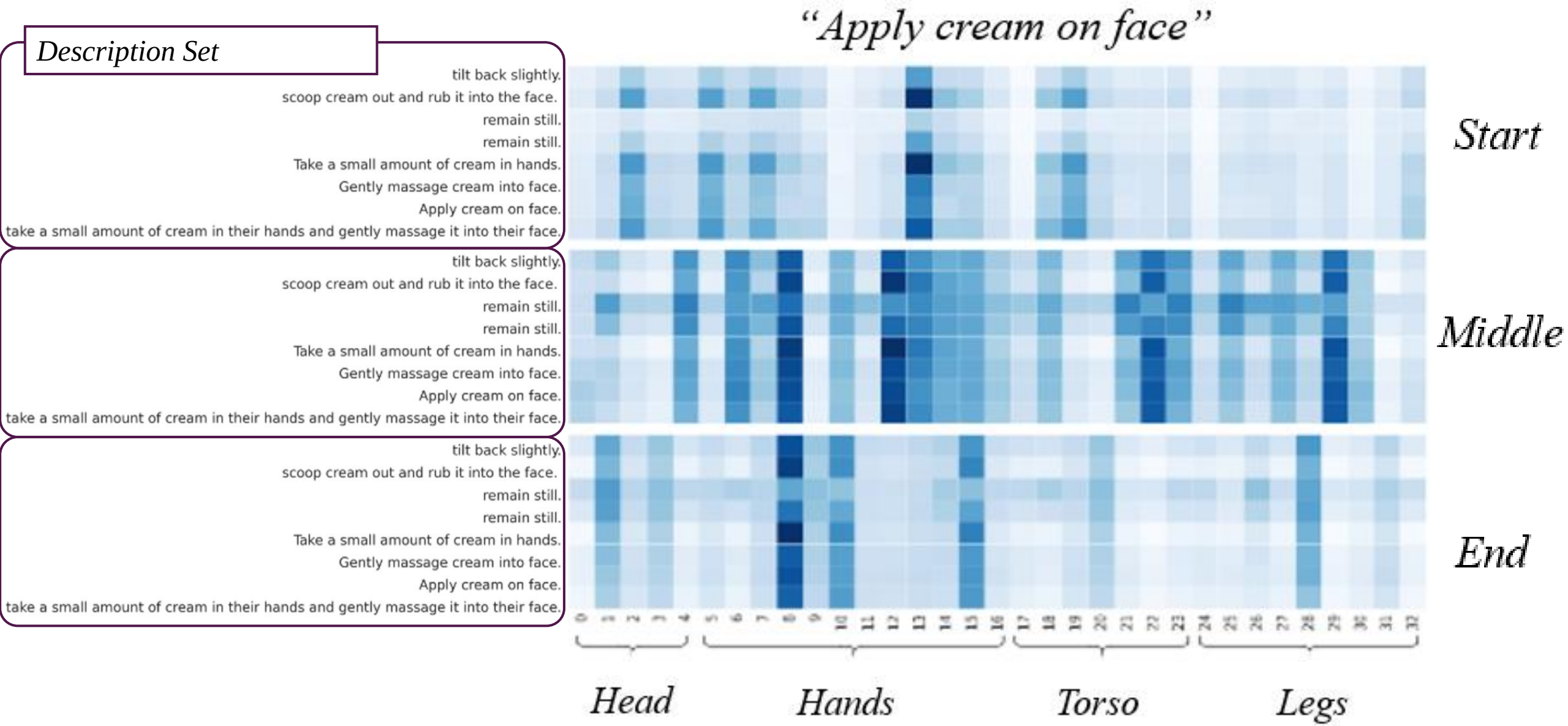
Partitioning Strategy	NTU-RGBD 60 (Acc %)				NTU-RGBD 120 (Acc %)			
	55/5	48/12	40/20	30/30	110/10	96/24	80/40	60/60
Global (Original)	64.69	35.46	27.15	16.29	66.96	44.27	21.31	14.12
Global (GPT-3)	78.50	33.47	29.21	22.27	64.89	47.15	25.16	17.46
Static	76.46	33.03	29.57	22.00	67.62	46.83	26.98	18.03
Adaptive	**79.23**	**40.99**	**31.05**	**23.52**	**71.95**	**52.01**	**28.38**	**19.63**

(c) Partitioning Strategy

α_i	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
-			78.50	33.47	29.21	22.27	64.89	47.15	25.16	17.46
Average	✓		76.68	37.80	30.92	22.20	68.11	30.93	24.36	18.67
Learnable	✓		76.32	37.62	29.06	21.91	71.73	40.92	23.49	19.13
Average		✓	78.65	38.80	28.14	22.69	55.73	50.67	27.50	17.50
Learnable		✓	77.70	40.69	28.84	22.46	71.26	46.13	24.43	18.57
Average	✓	✓	79.02	39.92	31.00	23.47	**73.55**	51.38	27.67	18.66
Learnable	✓	✓	**79.23**	**40.99**	**31.05**	**23.52**	71.95	**52.01**	**28.38**	**19.63**

(d) Local Knowledge Transfer

Visualization of Adaptive Partitioning



Thank you for watching.