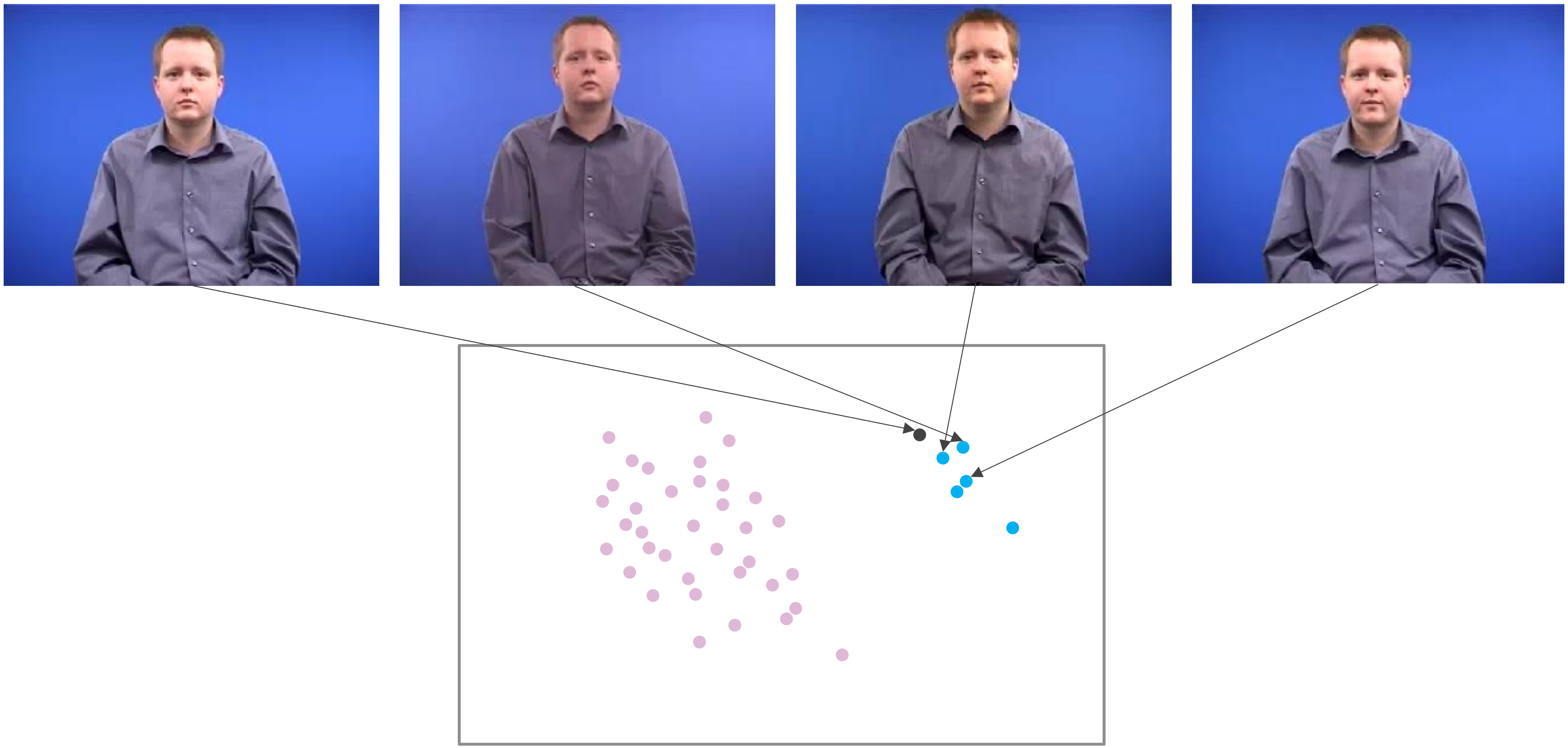
Geo-Sign: Hyperbolic Contrastive Regularisation for Geometrically Aware Sign Language Translation

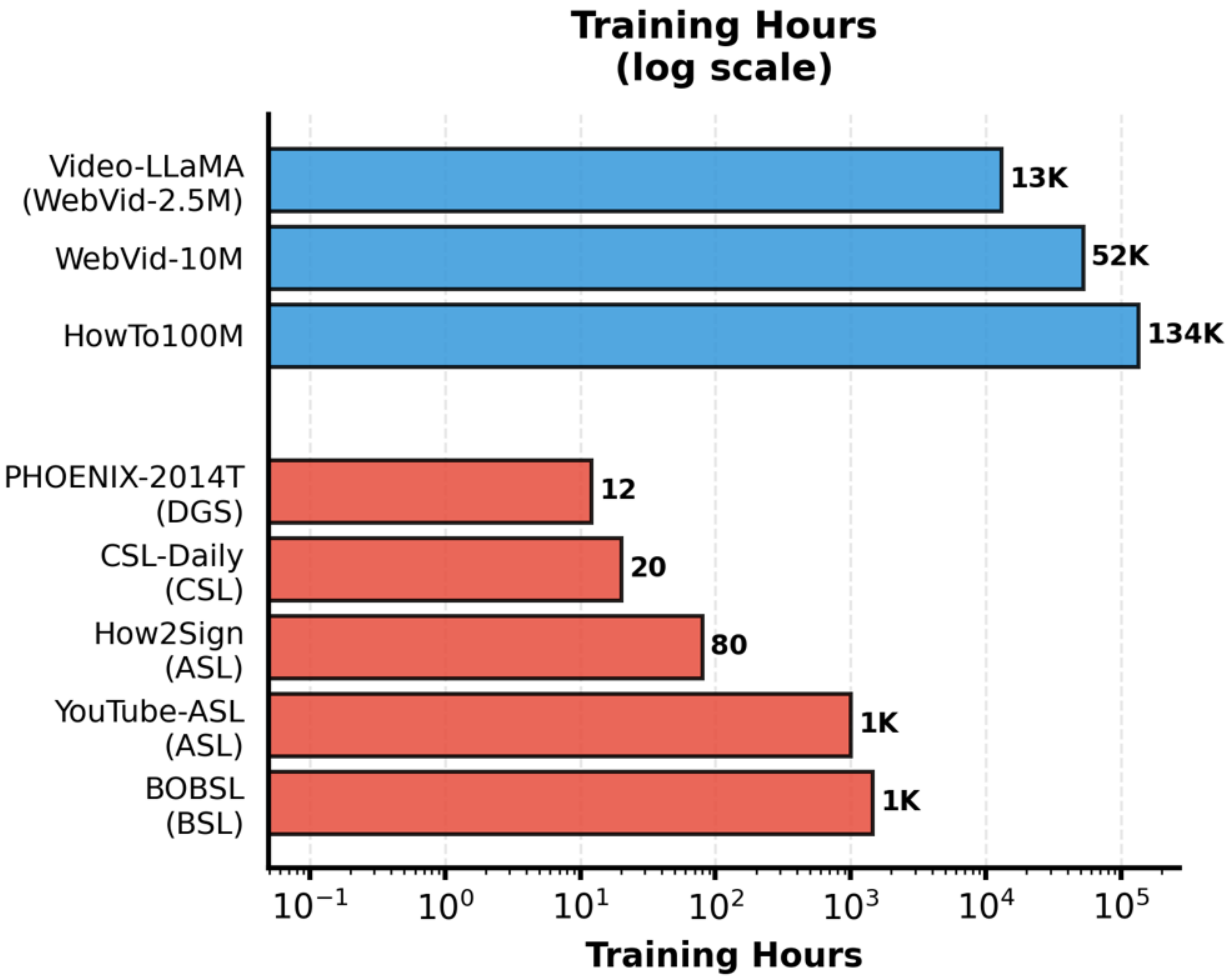
Edward Fish & Richard Bowden

University of Surrey, UK

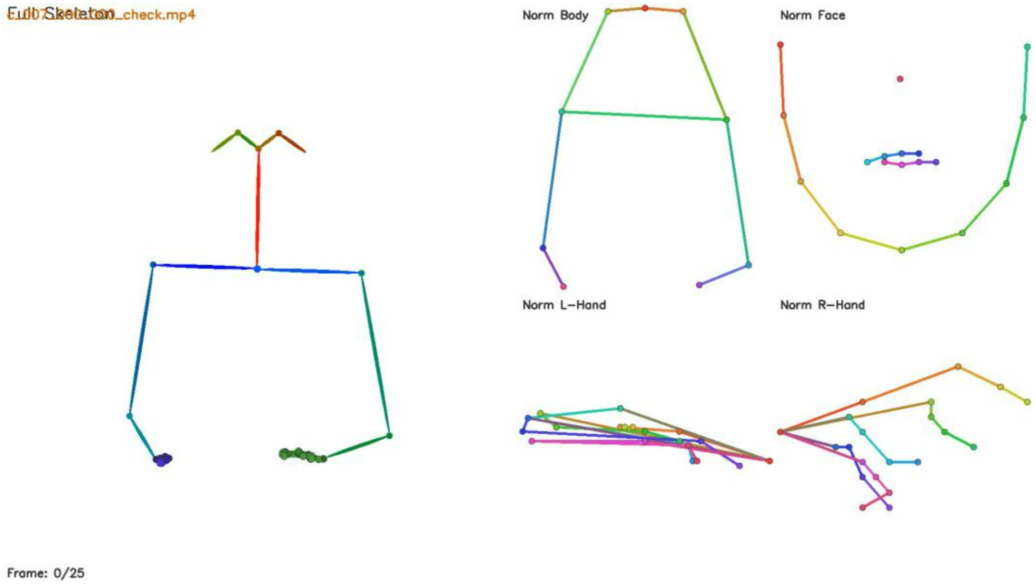
NEURAL INFORMATION PROCESSING SYSTEMS

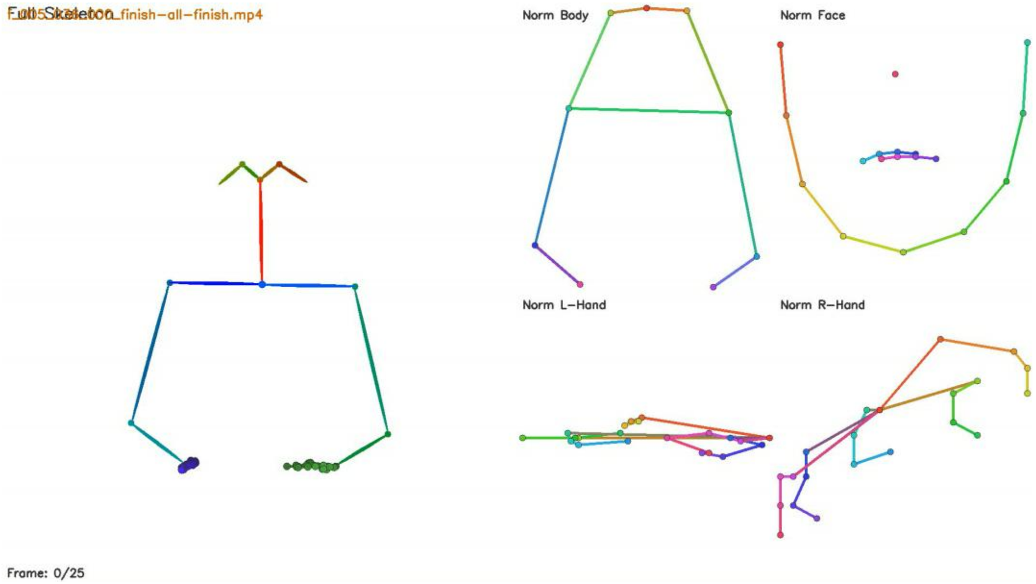
CVSSP Centre for Vision, Speech and Signal Processing

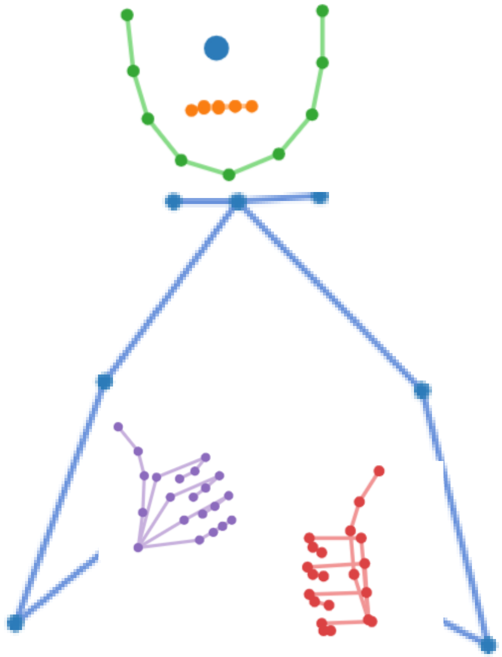


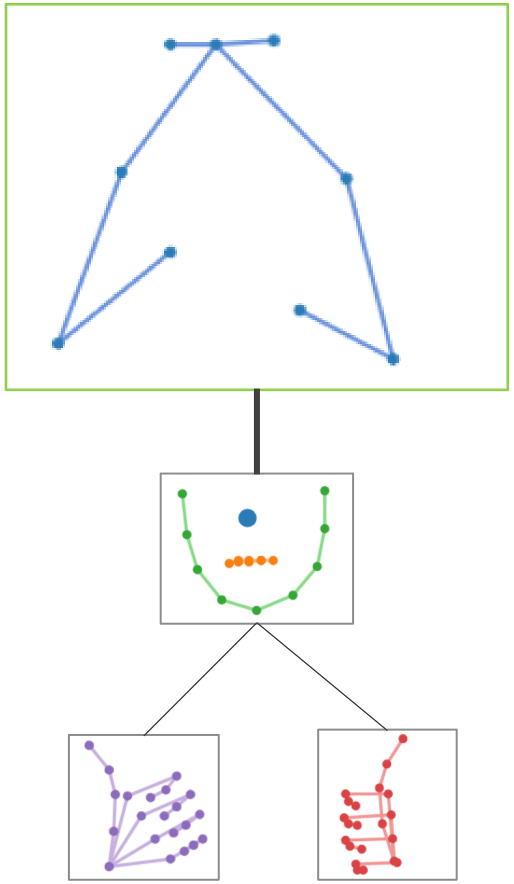

Training Hours
(log scale)
Video-LLaMA
(WebVid-2.5M)
13K
WebVid-10M
52K
HowTo100M
134K
PHOENIX-2014T
(DGS)
12
CSL-Daily
(CSL)
20
How2Sign
(ASL)
80
YouTube-ASL
(ASL)
1K
BOBSL
(BSL)
1K
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}
Training Hours

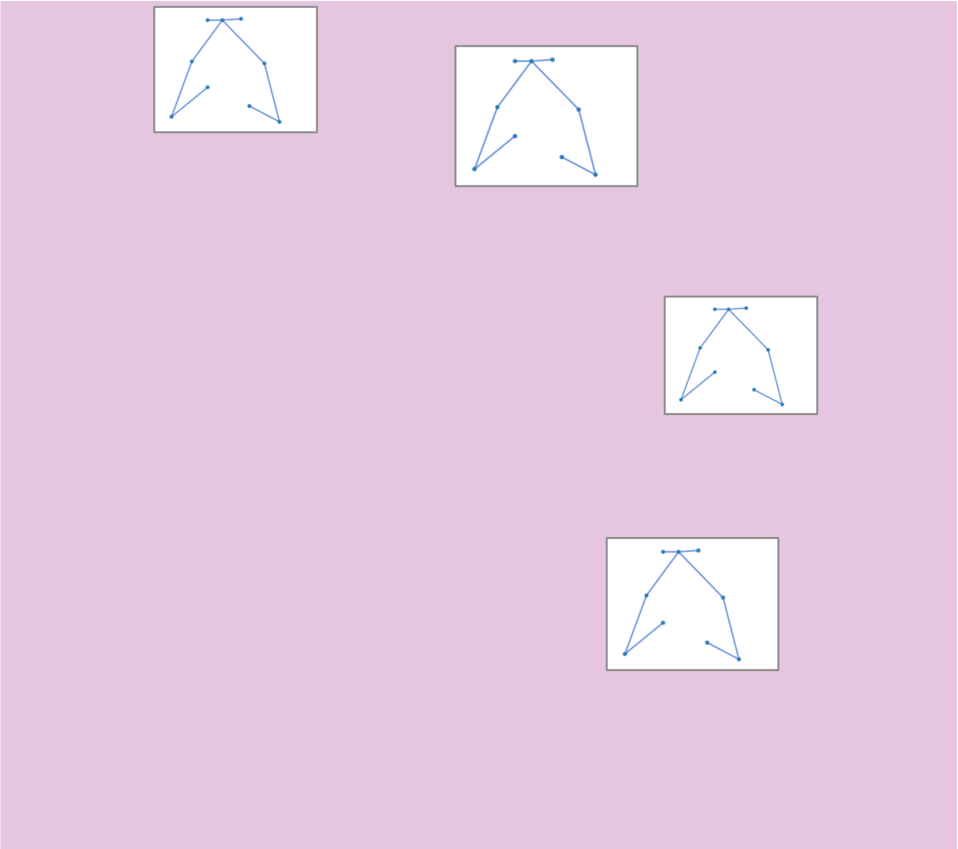
How can we create better representations without additional data?

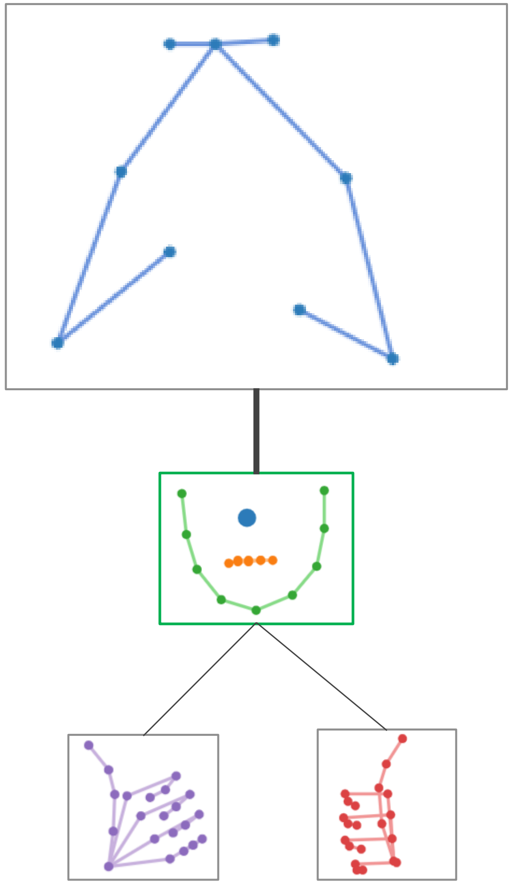

Full Skeleton
Norm Body
Norm Face
Norm L-Hand
Norm R-Hand
Frame: 0/25

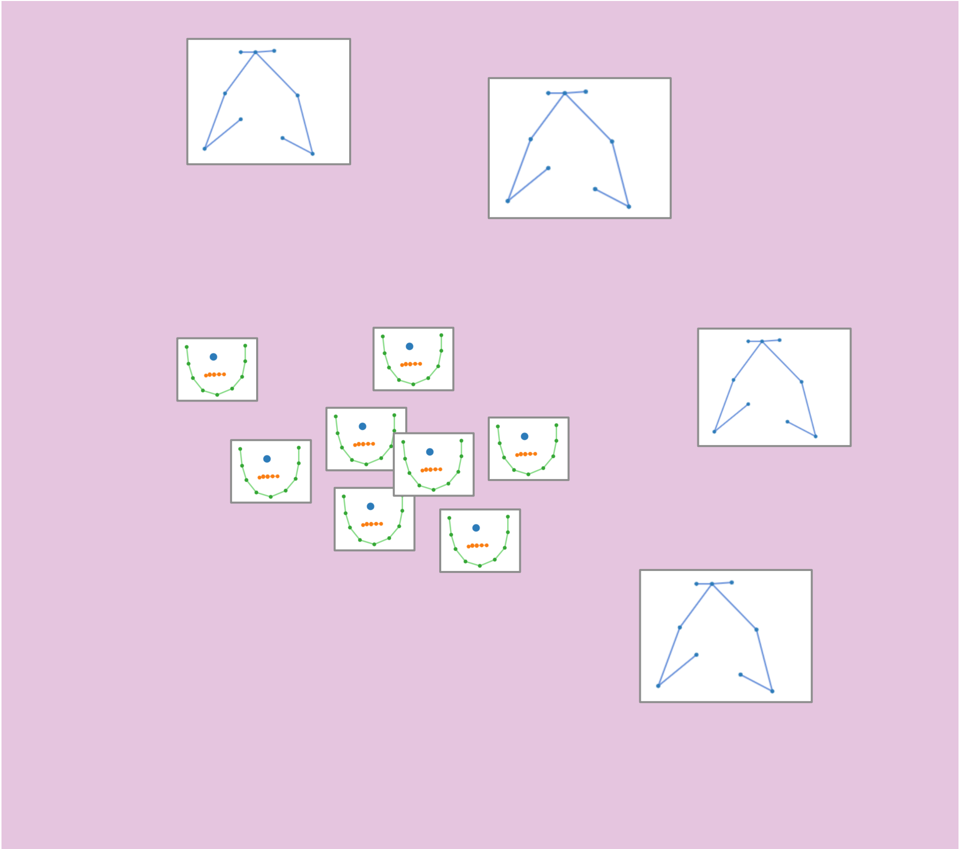

Full Skeleton
Norm Body
Norm Face
Norm L-Hand
Norm R-Hand
Frame: 0/25

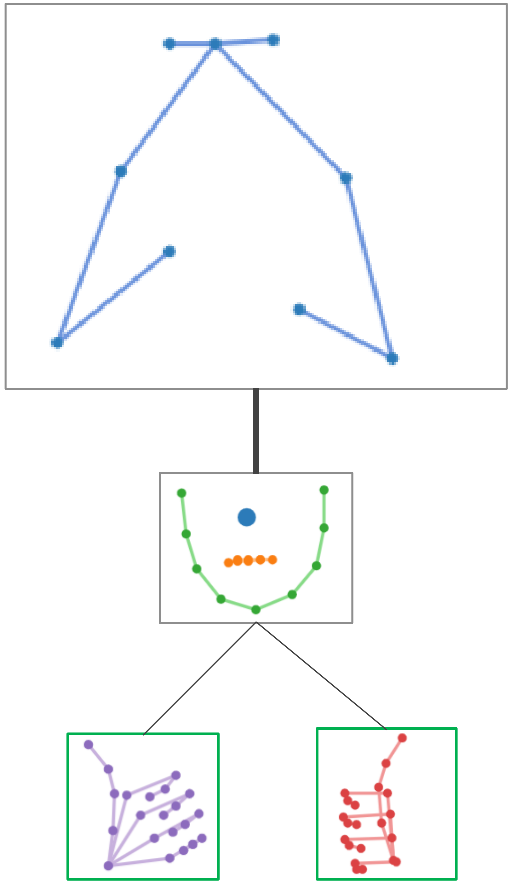


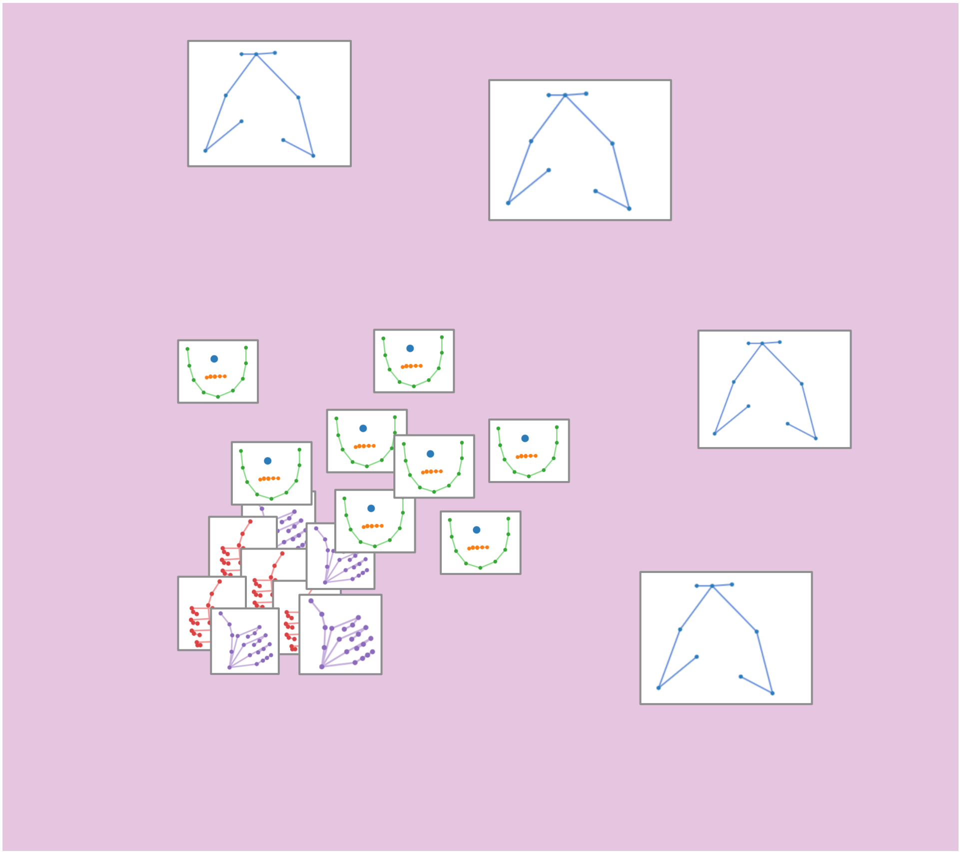


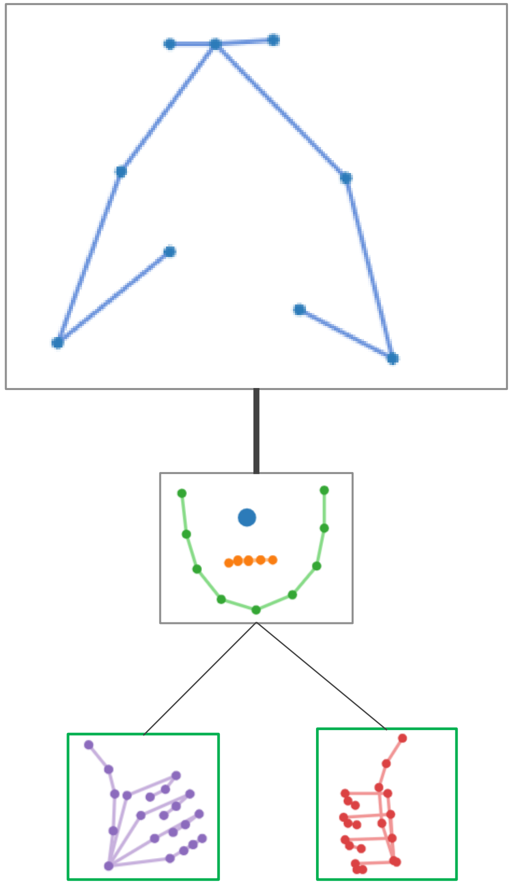


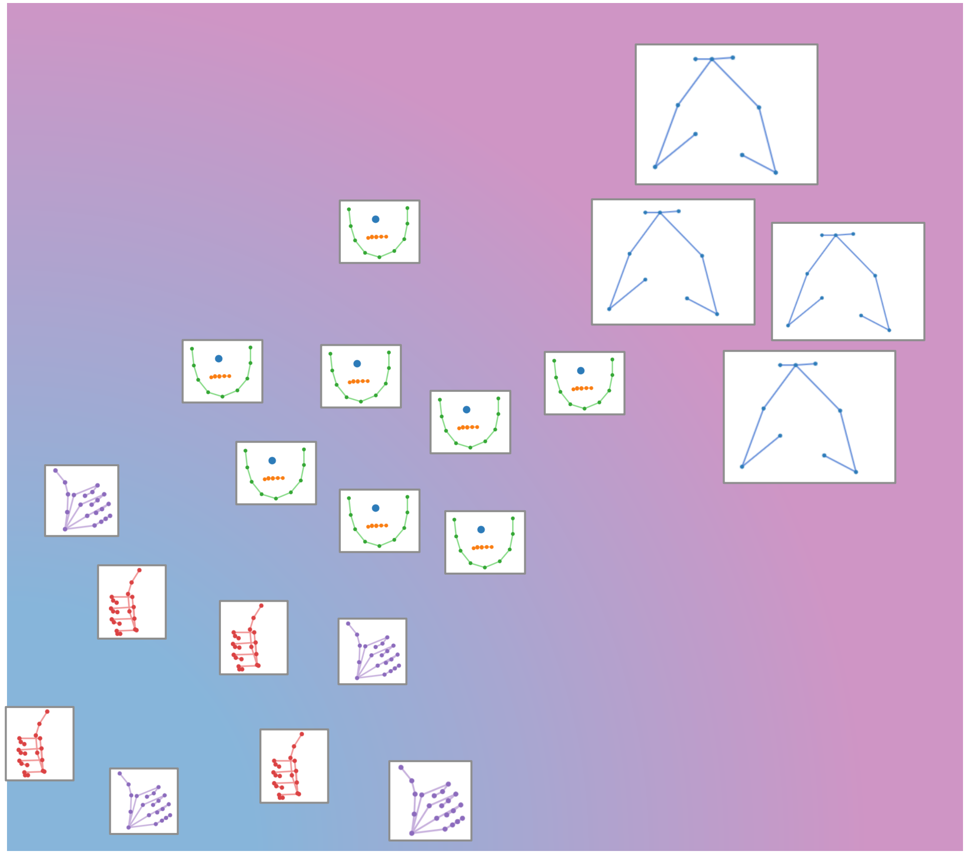




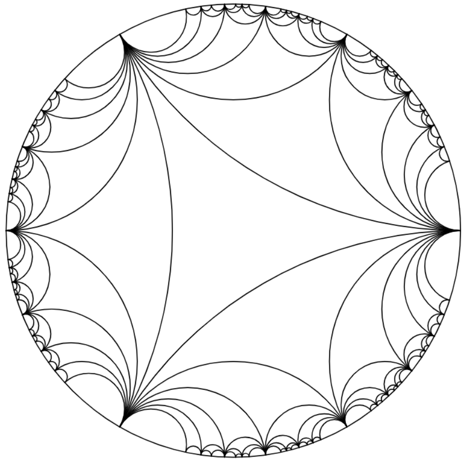


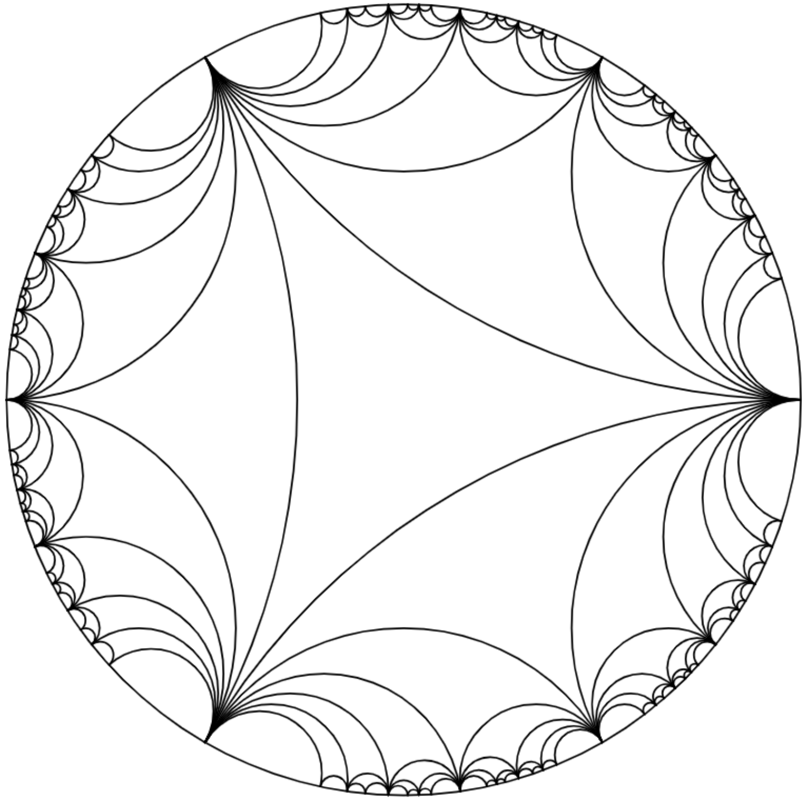


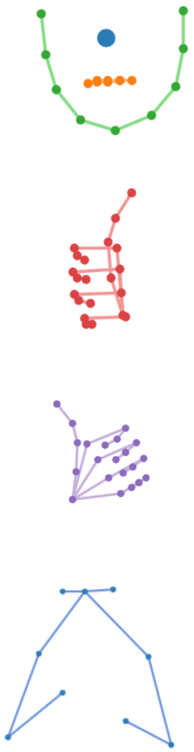


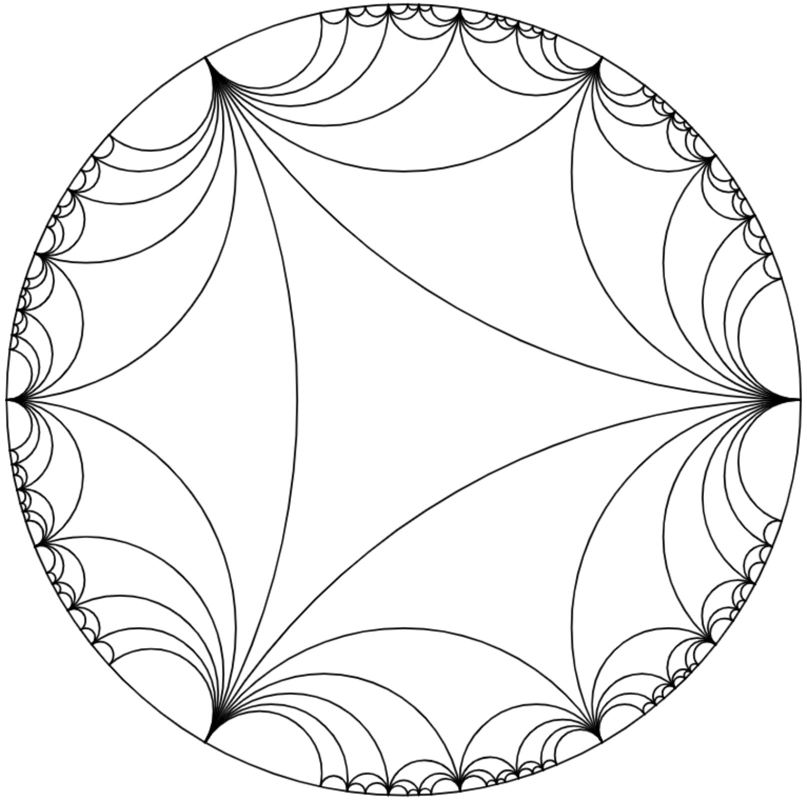


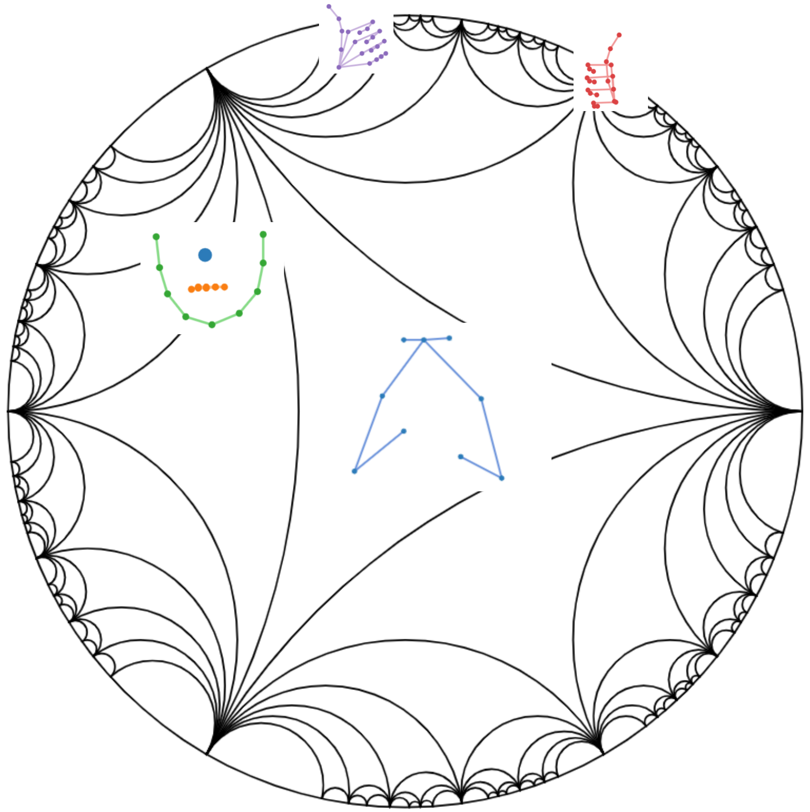
Hyperbolic Geometry

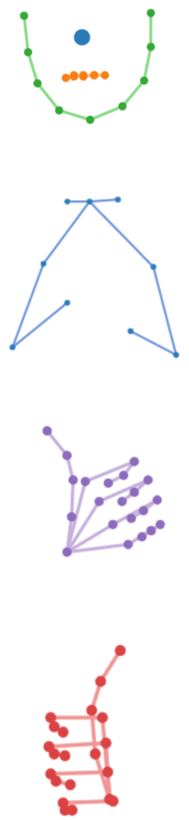


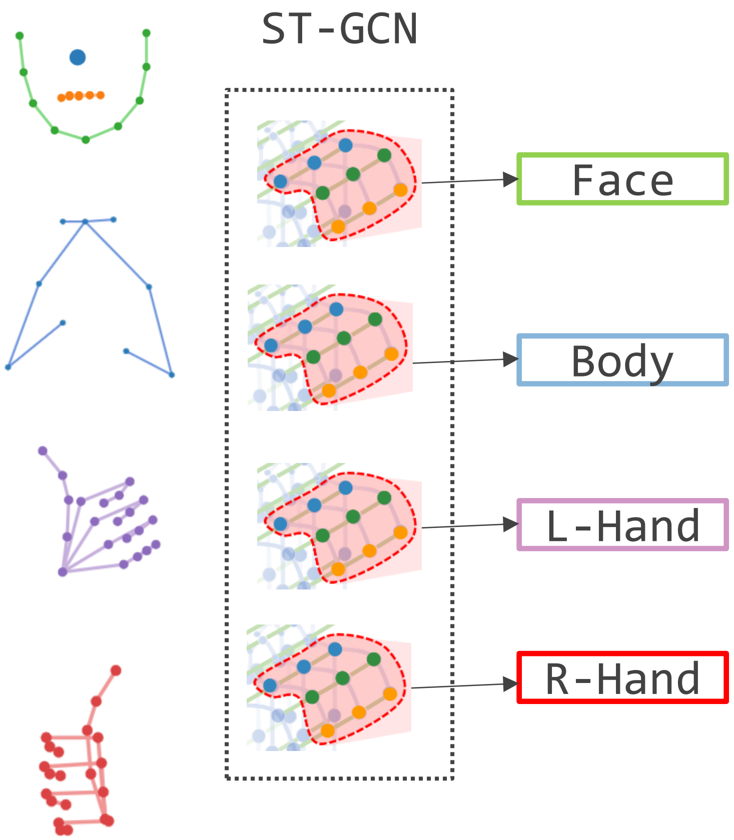




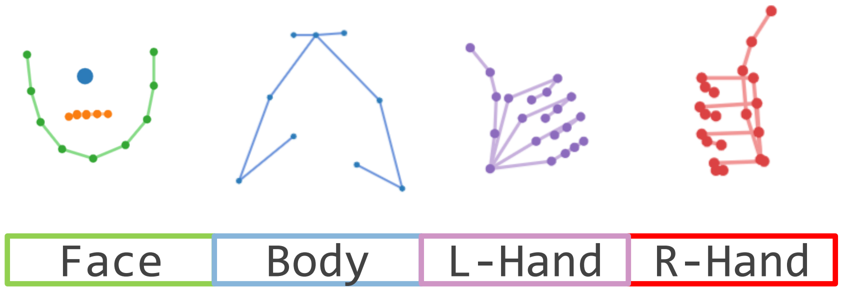


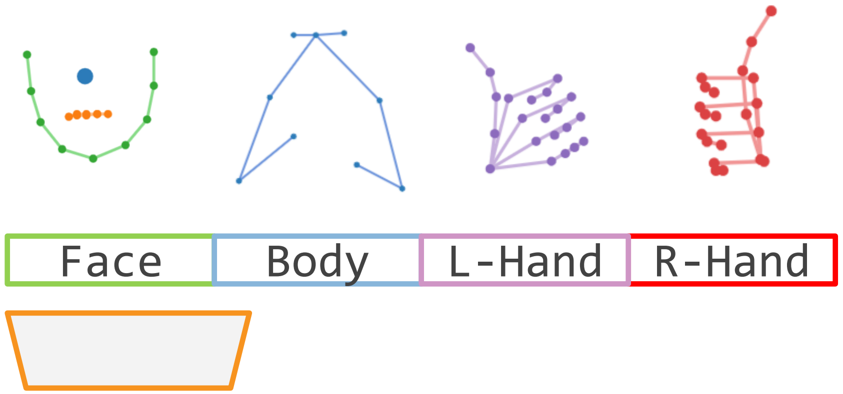


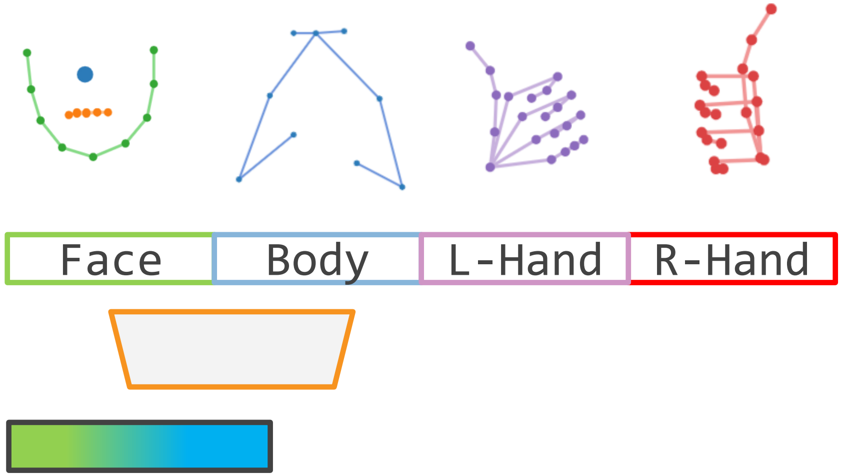


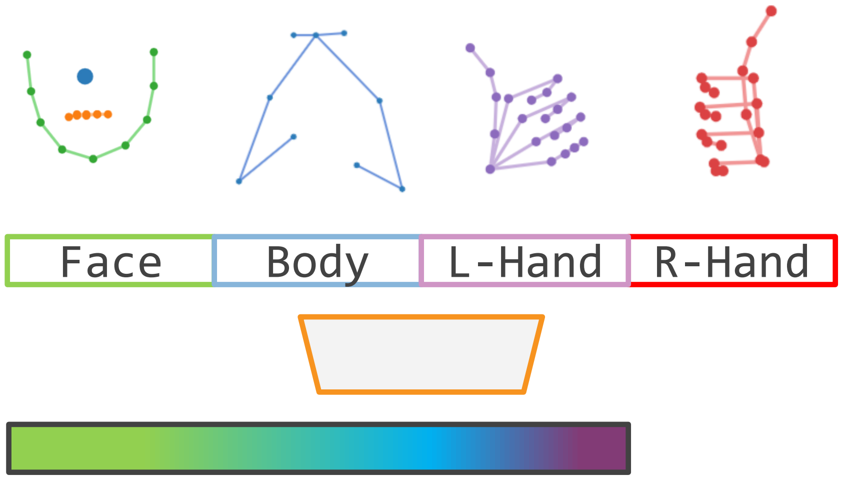


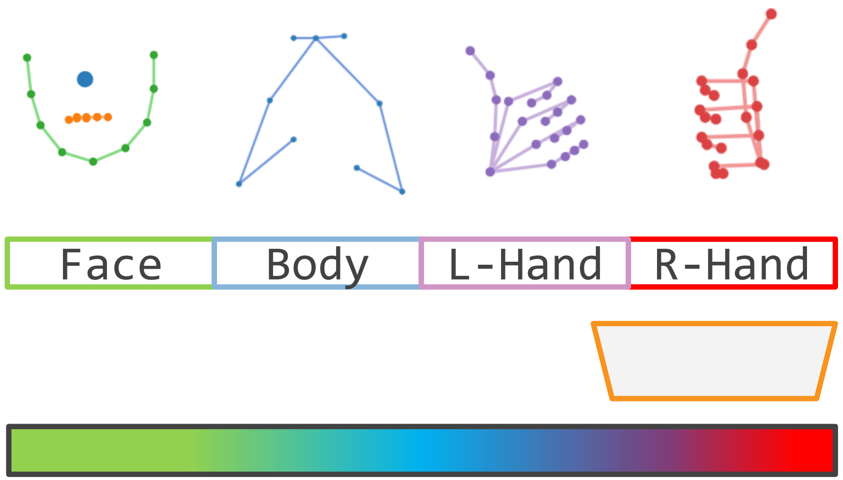
Yan, Sijie, Yuanjun Xiong, and Dahua Lin. "Spatial temporal graph convolutional networks for skeleton-based action recognition." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 32. No. 1. 2018.


Face
Body
L-Hand
R-Hand


Face
Body
L-Hand
R-Hand

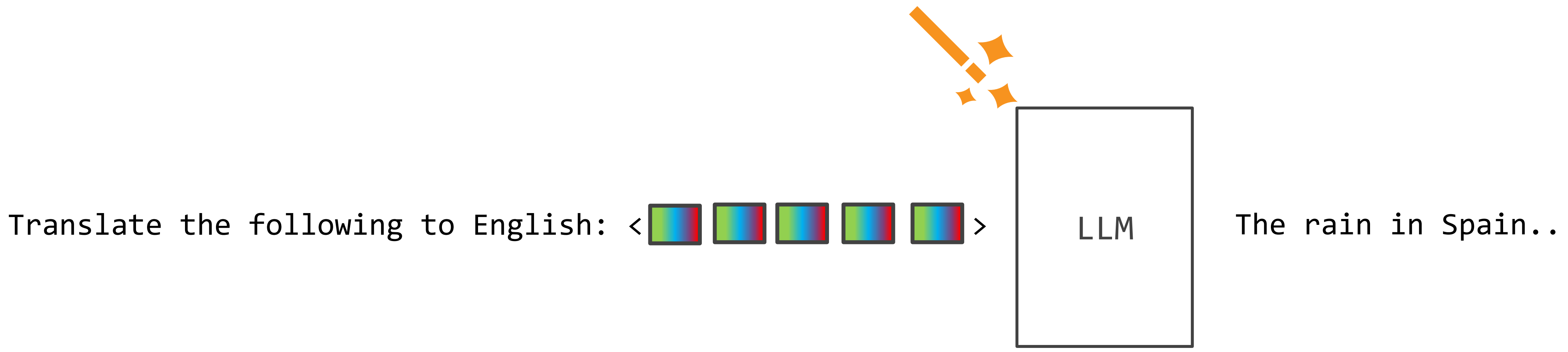

Face
Body
L-Hand
R-Hand

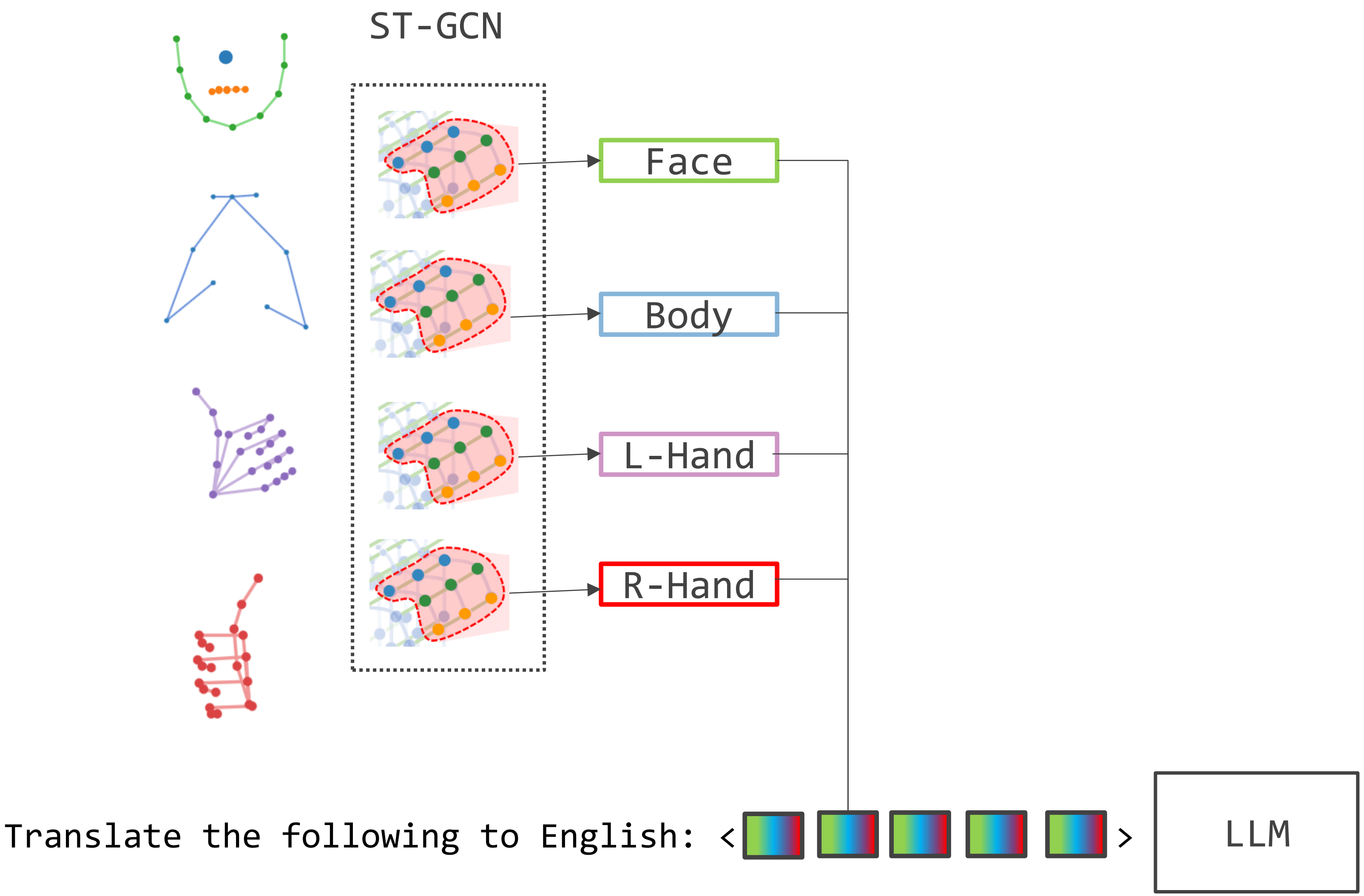

Face
Body
L-Hand
R-Hand

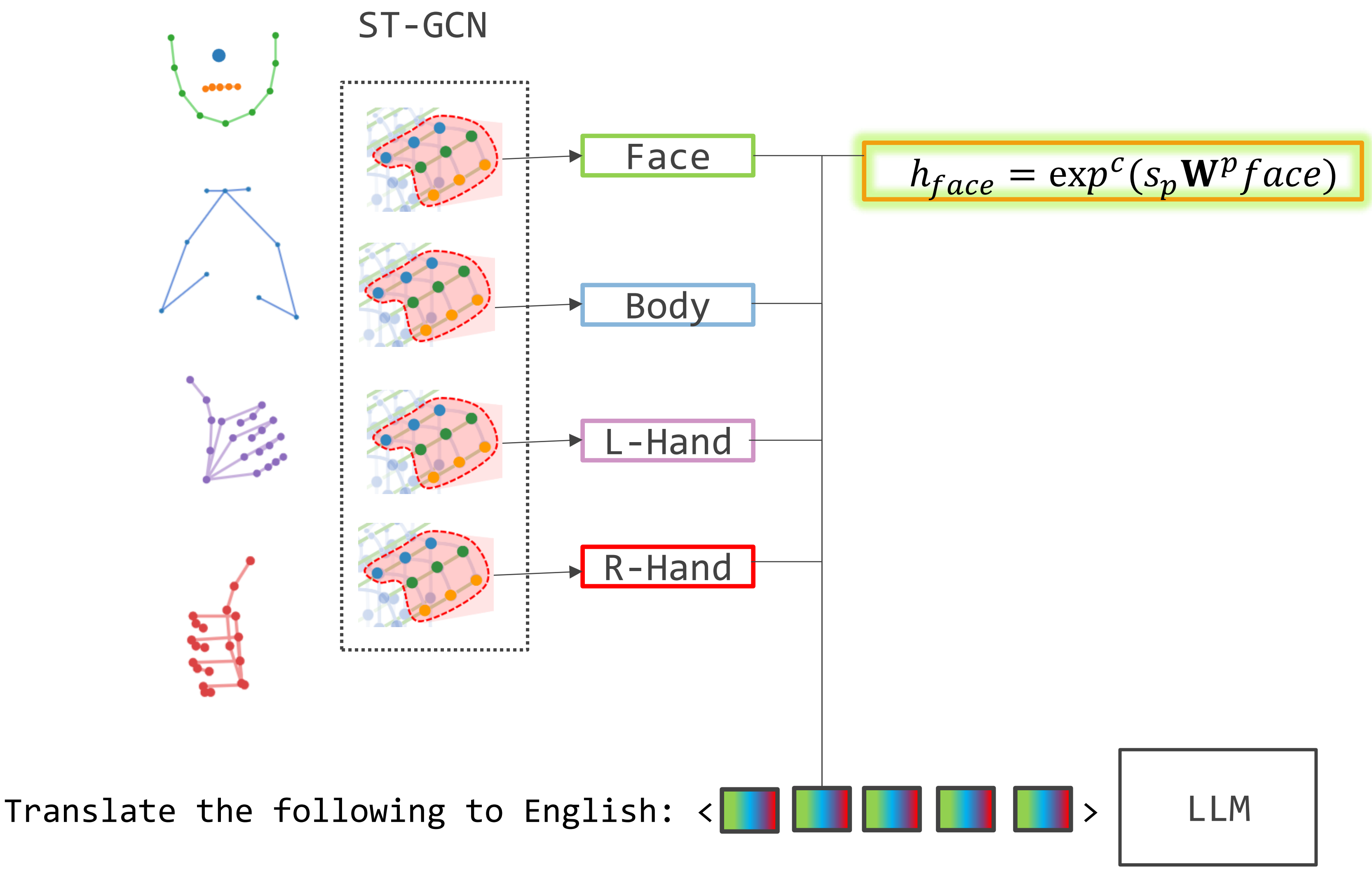

Face
Body
L-Hand
R-Hand

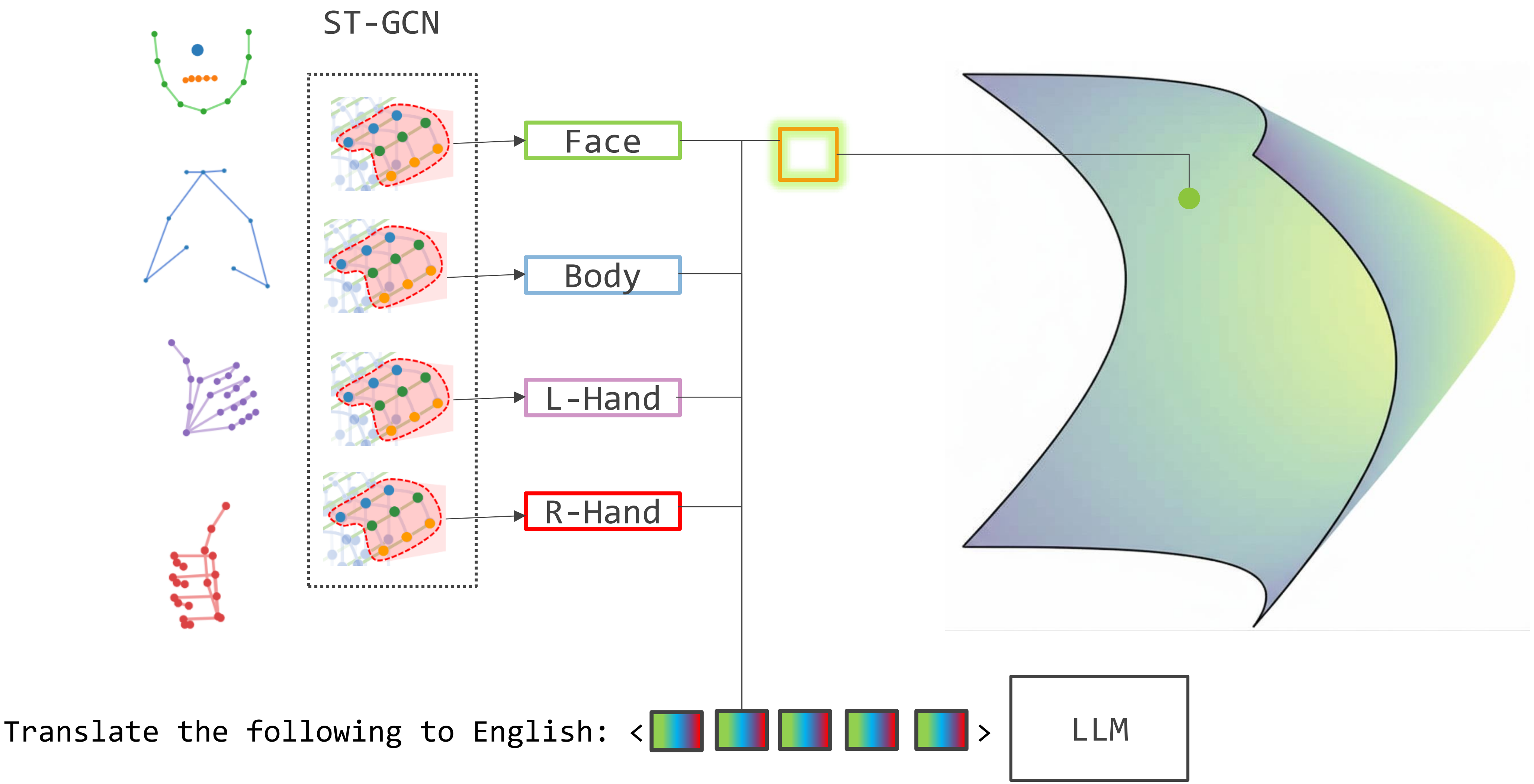
Translate the following to English: <  >

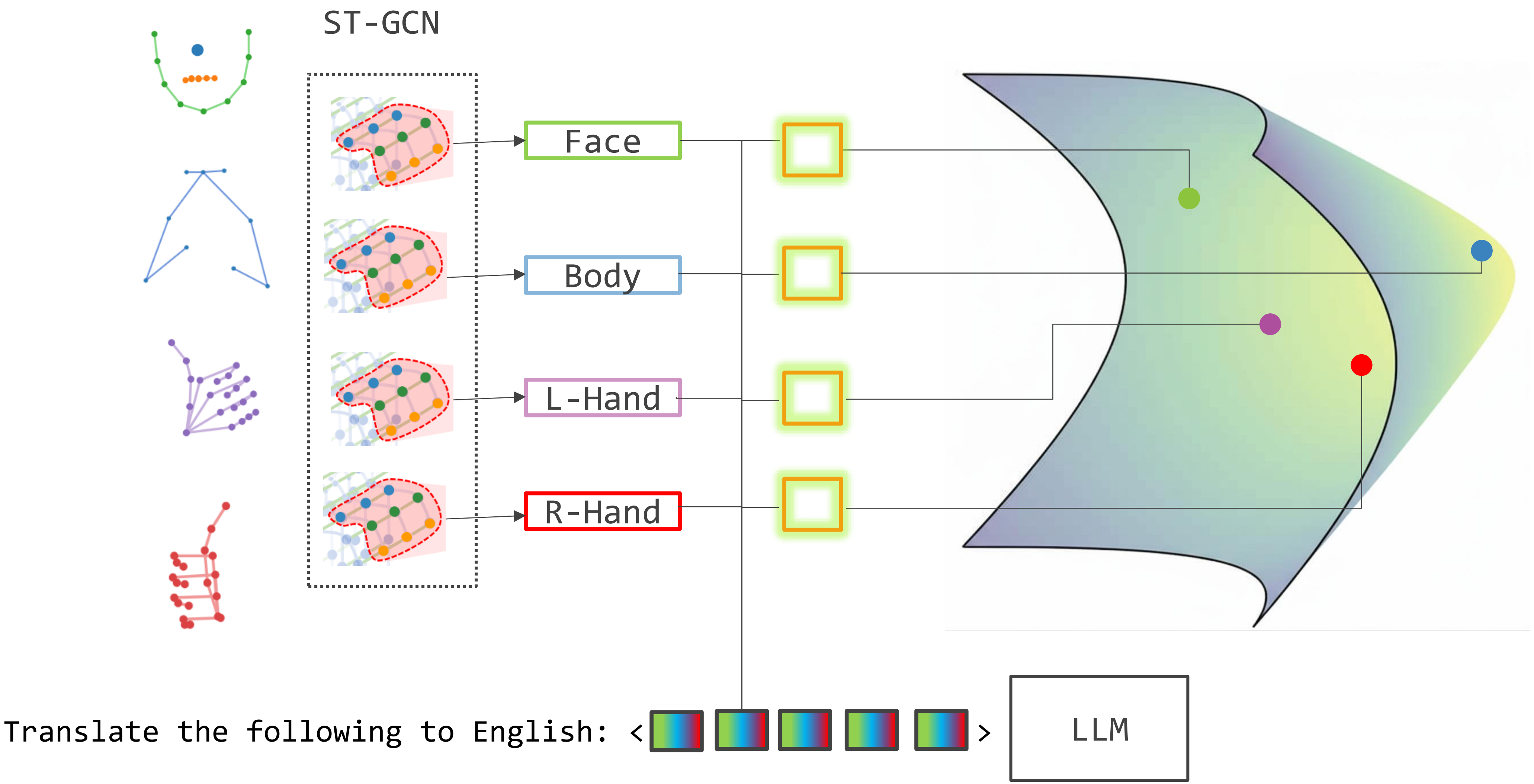

Translate the following to English: < >
LLM

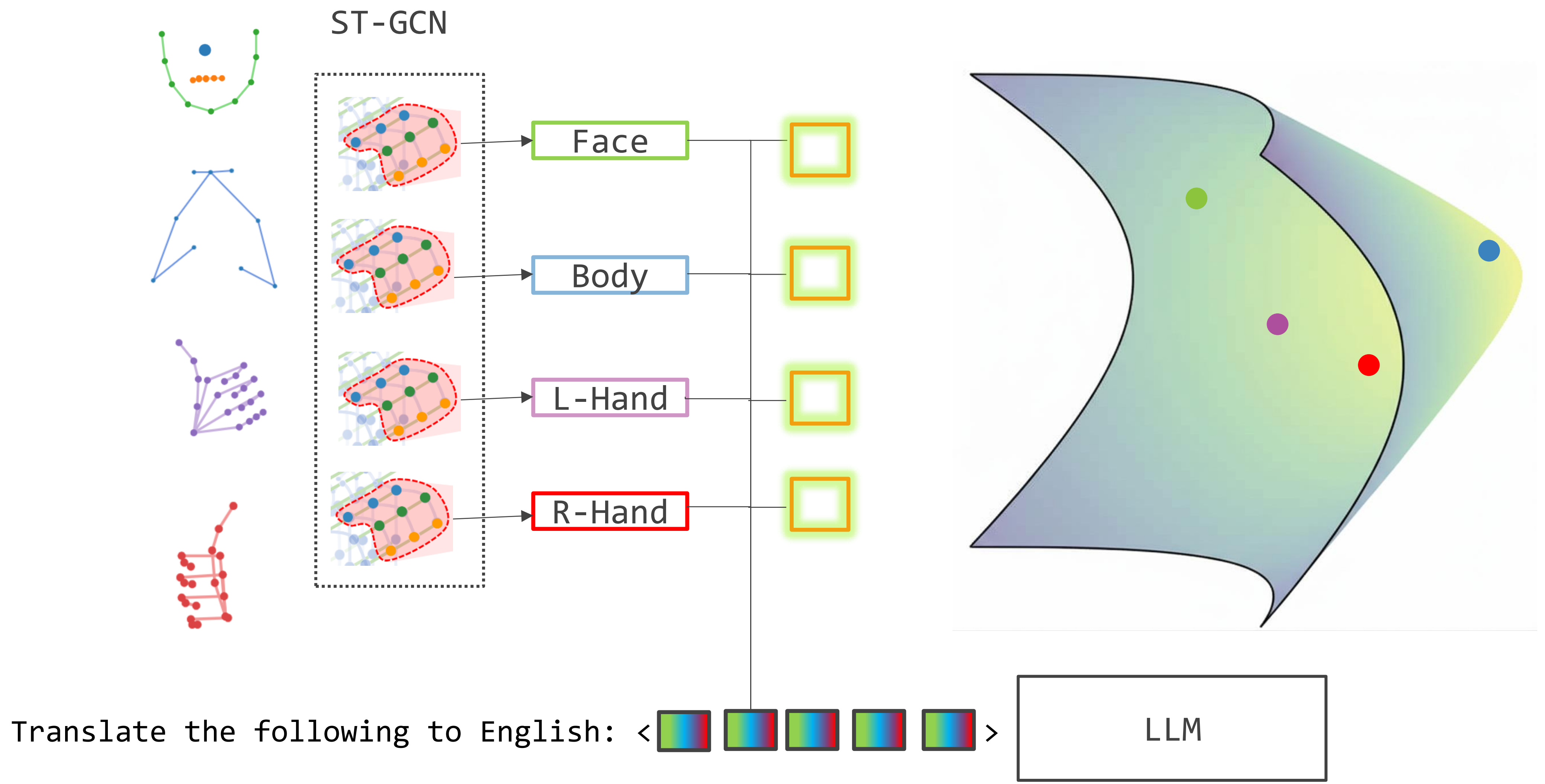

Translate the following to English: <
>
LLM
The rain in Spain..

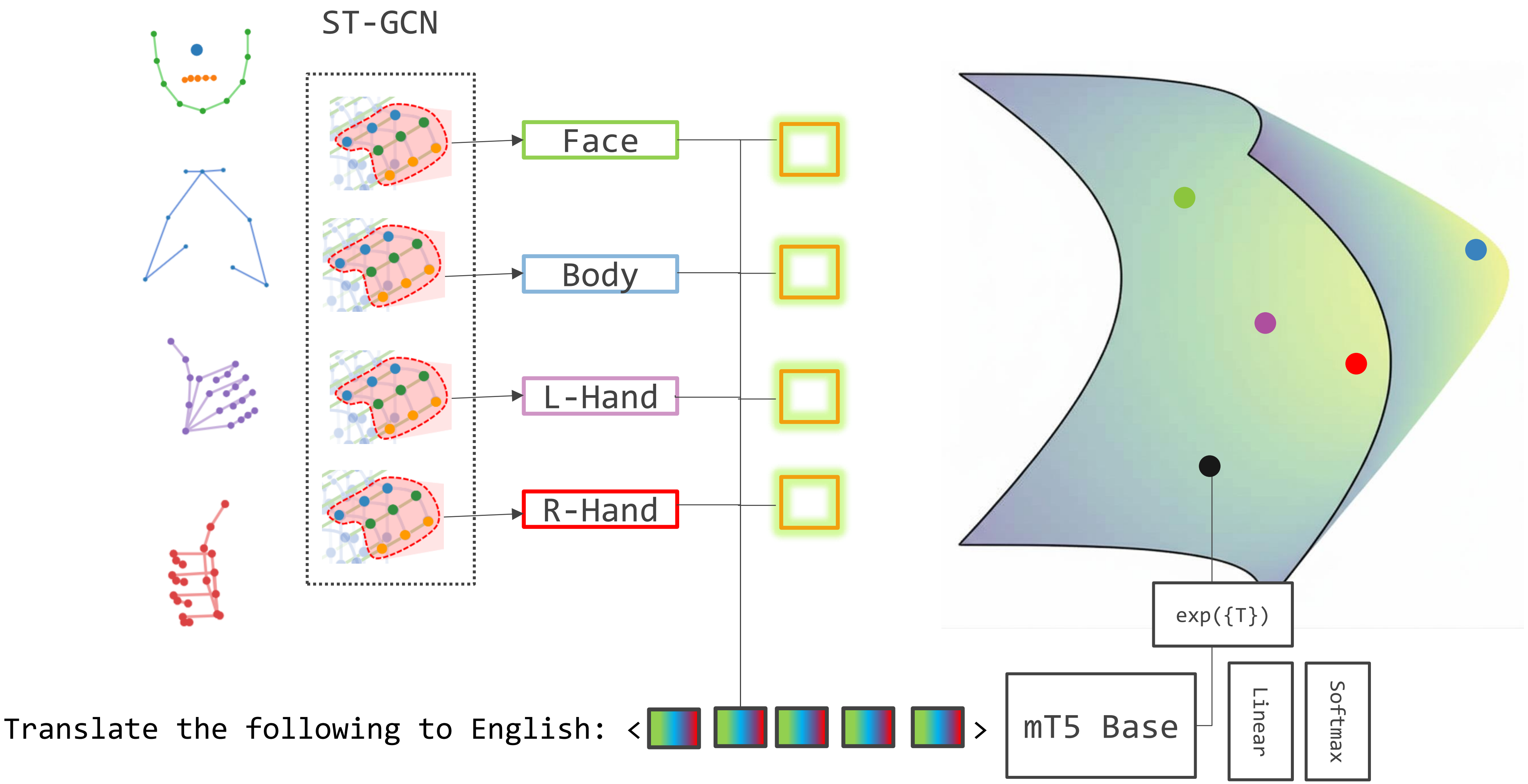

ST-GCN
Face
Body
L-Hand
R-Hand
Translate the following to English: <
>
LLM


ST-GCN
Face
Body
L-Hand
R-Hand
$h_{face} = \exp^c(s_p \mathbf{W}^p face)$
Translate the following to English: <
>
LLM


ST-GCN
Face
Body
L-Hand
R-Hand
Translate the following to English: < >
LLM

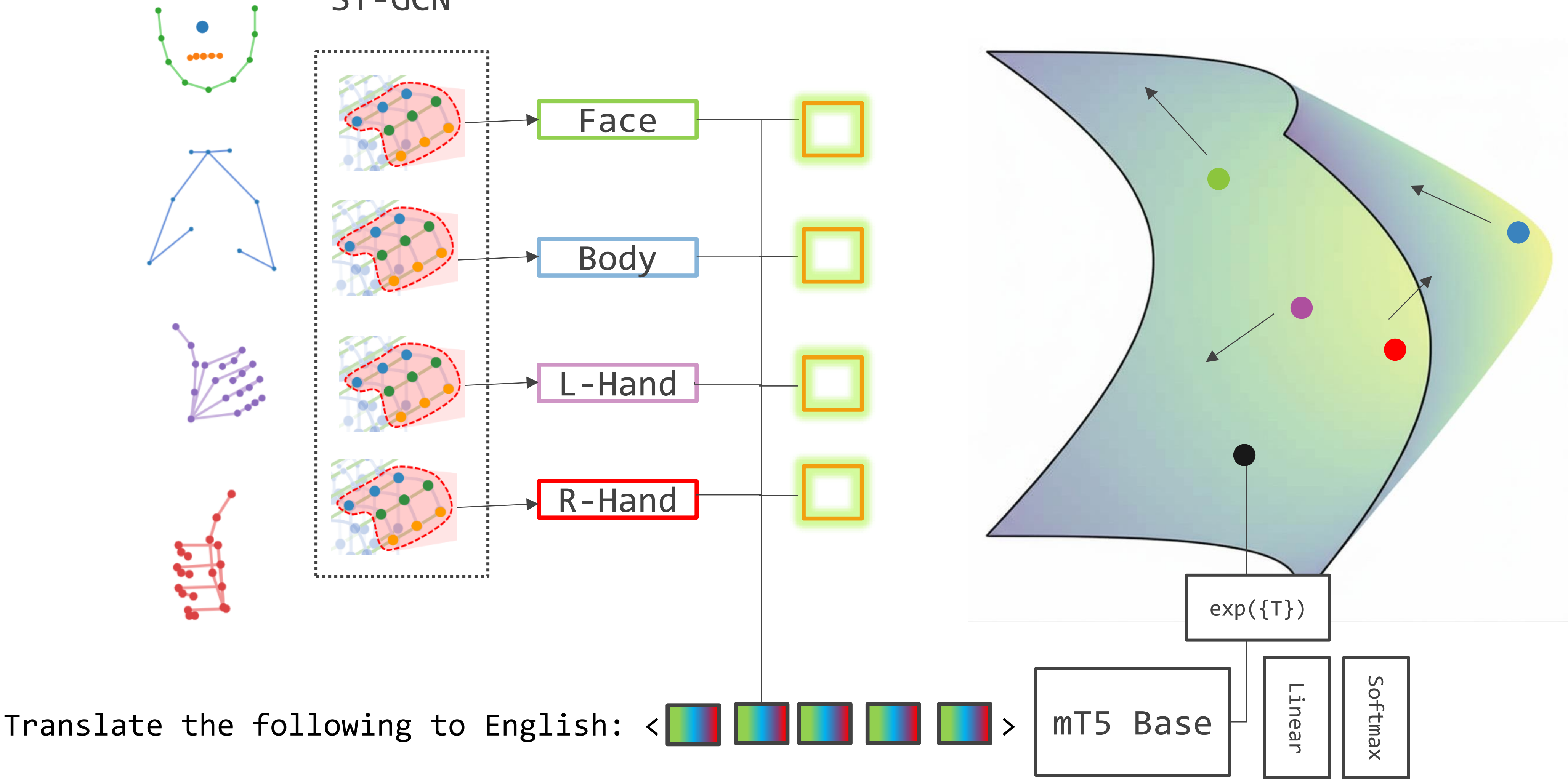

ST-GCN
Face
Body
L-Hand
R-Hand
Translate the following to English: <
>
LLM


ST-GCN
Face
Body
L-Hand
R-Hand
Translate the following to English: <
>
LLM

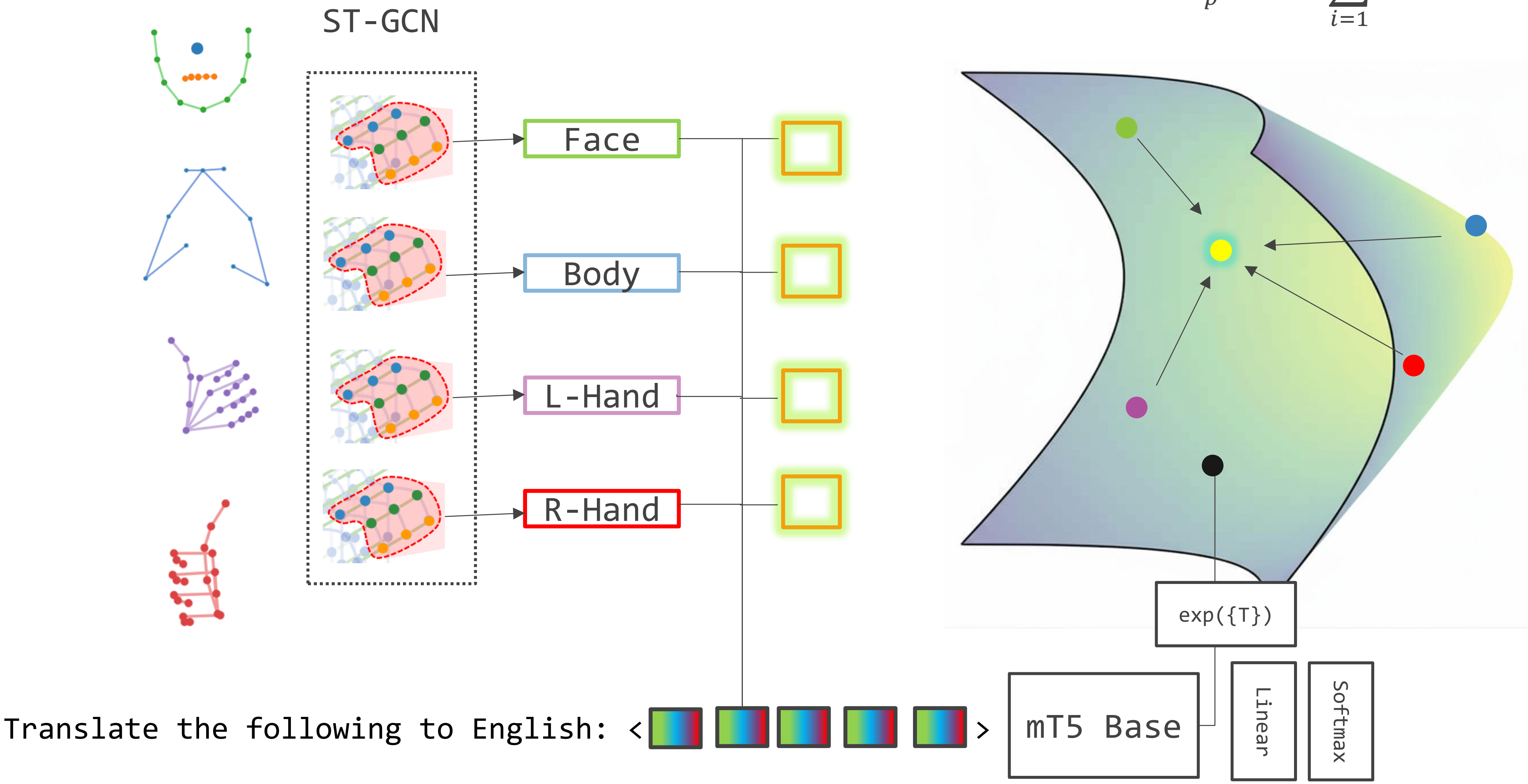


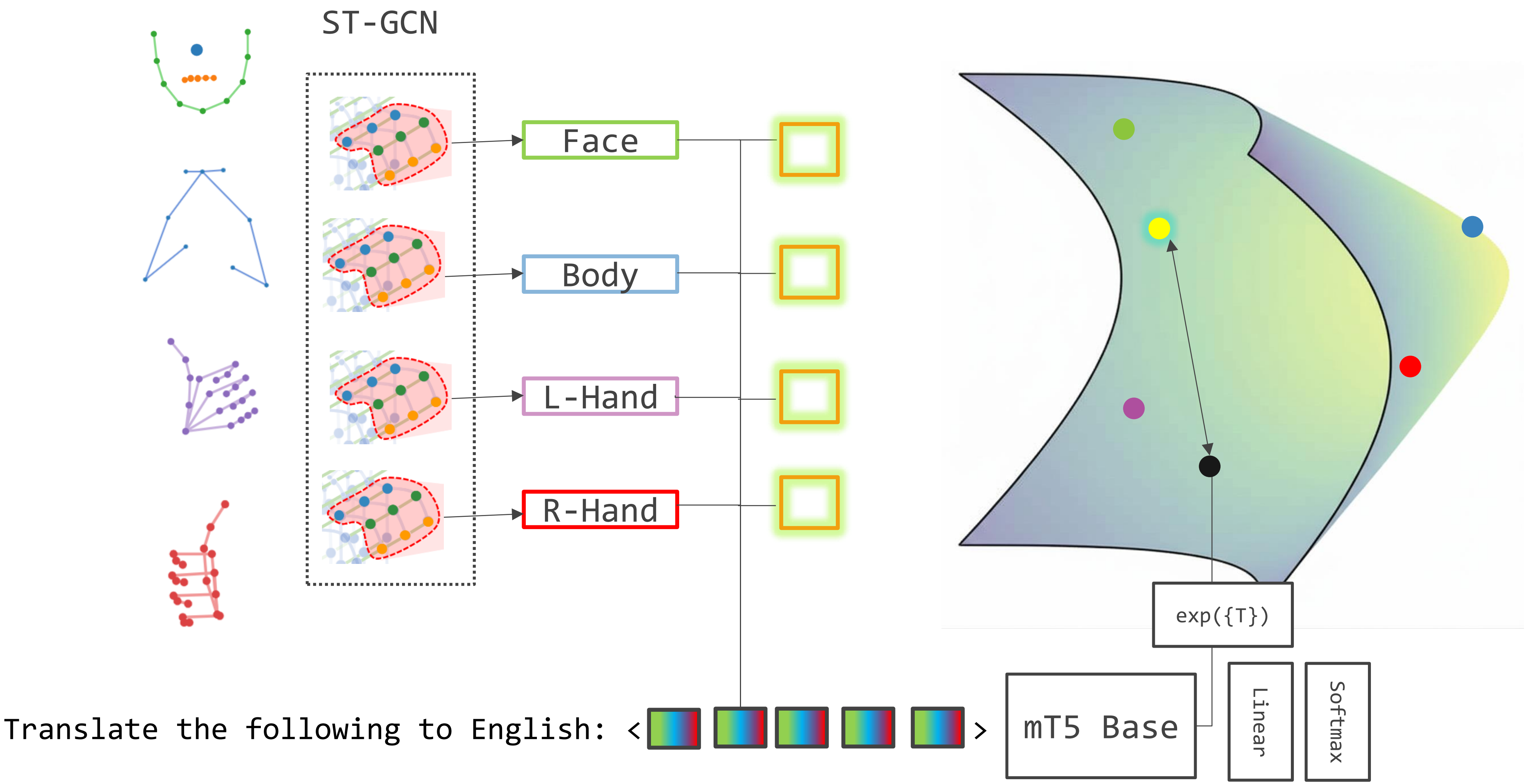
ST-GCN
Face
Body
L-Hand
R-Hand
exp({T})
Translate the following to English: <
>
mT5 Base
Linear
Softmax

$$h_z = \exp^c(s_p \mathbf{W}^p z)$$

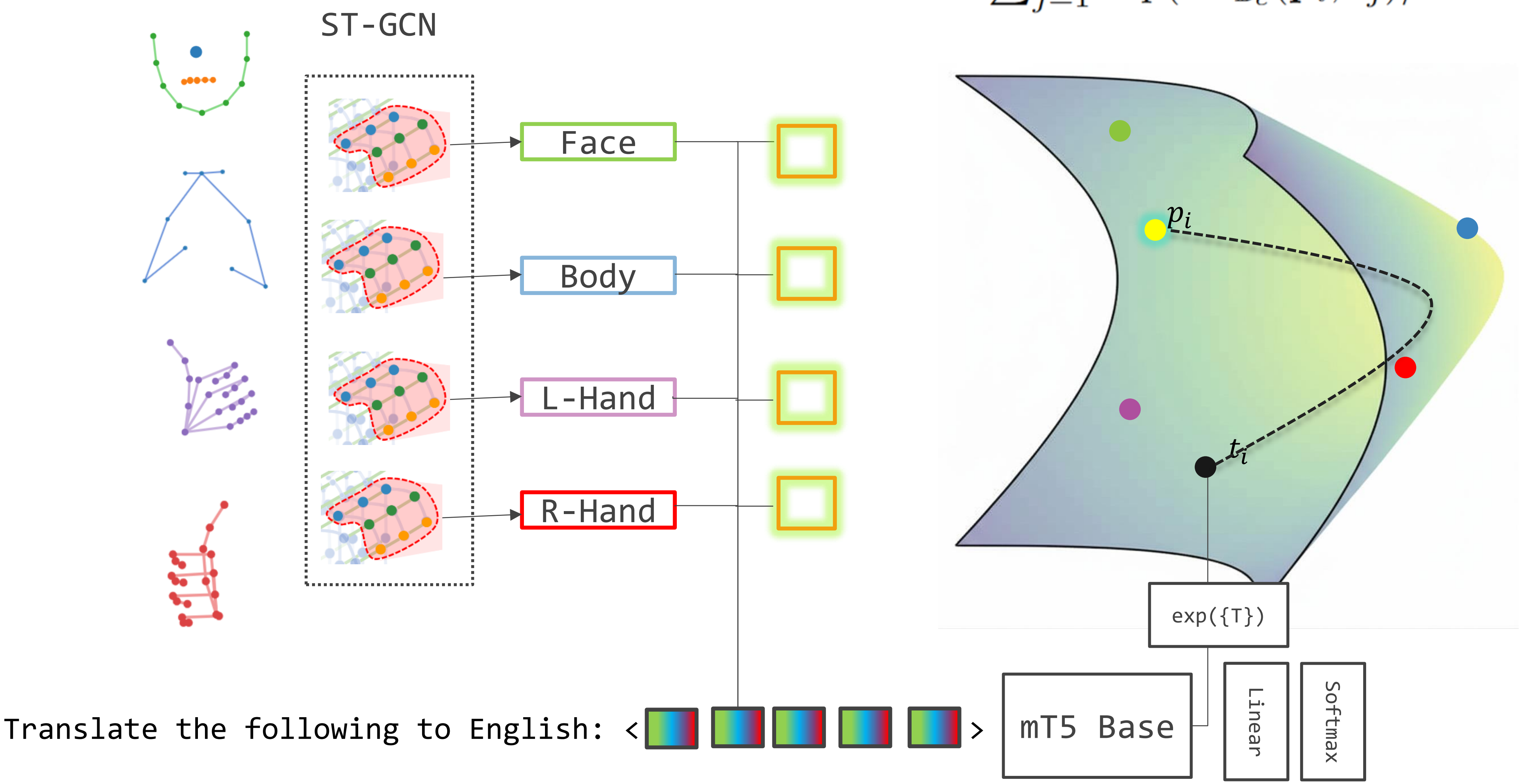


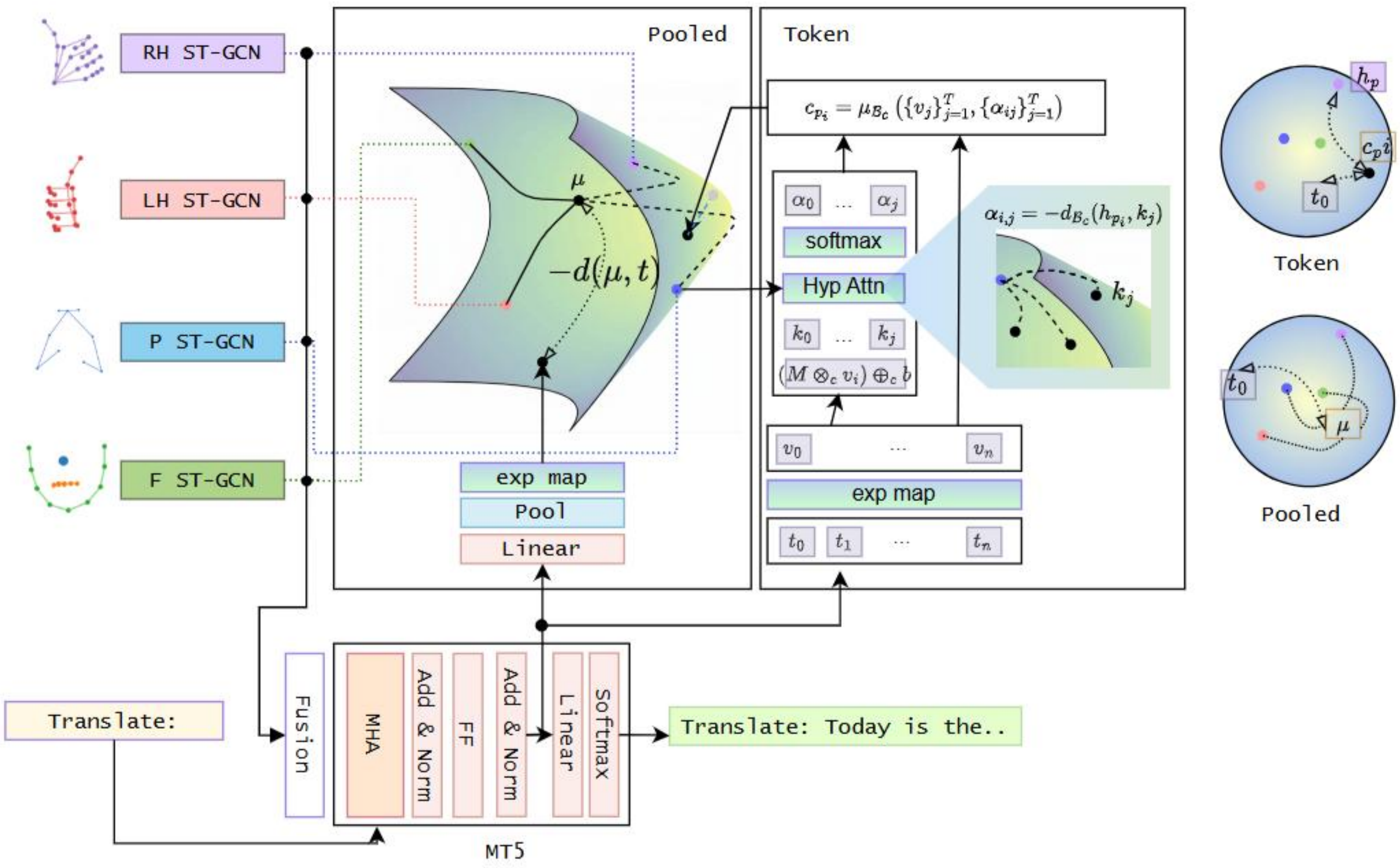
$$m = \underset{p}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \cdot d(x_i, p)^2$$



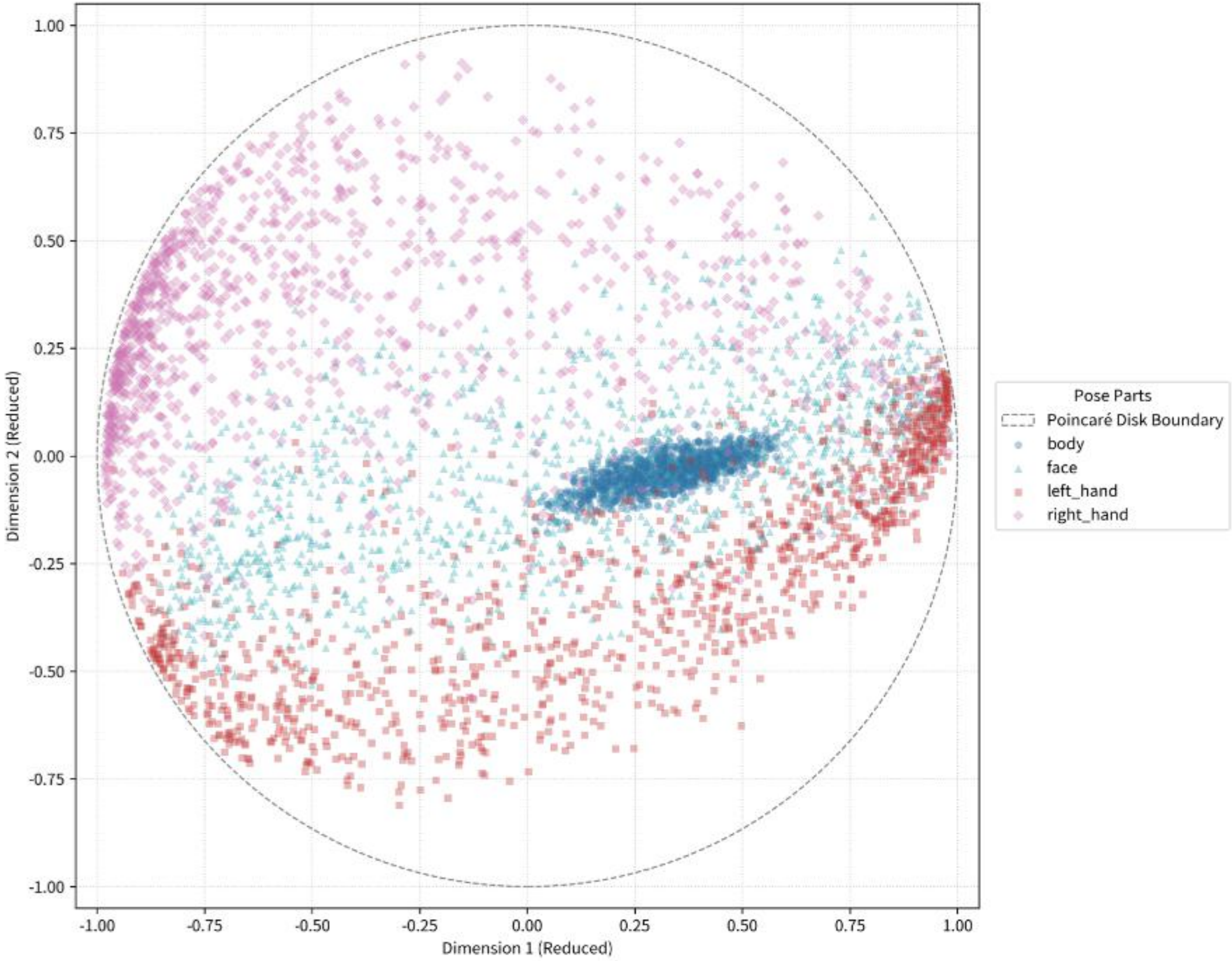

ST-GCN
Face
Body
L-Hand
R-Hand
exp({T})
mT5 Base
Linear
Softmax
Translate the following to English: < >

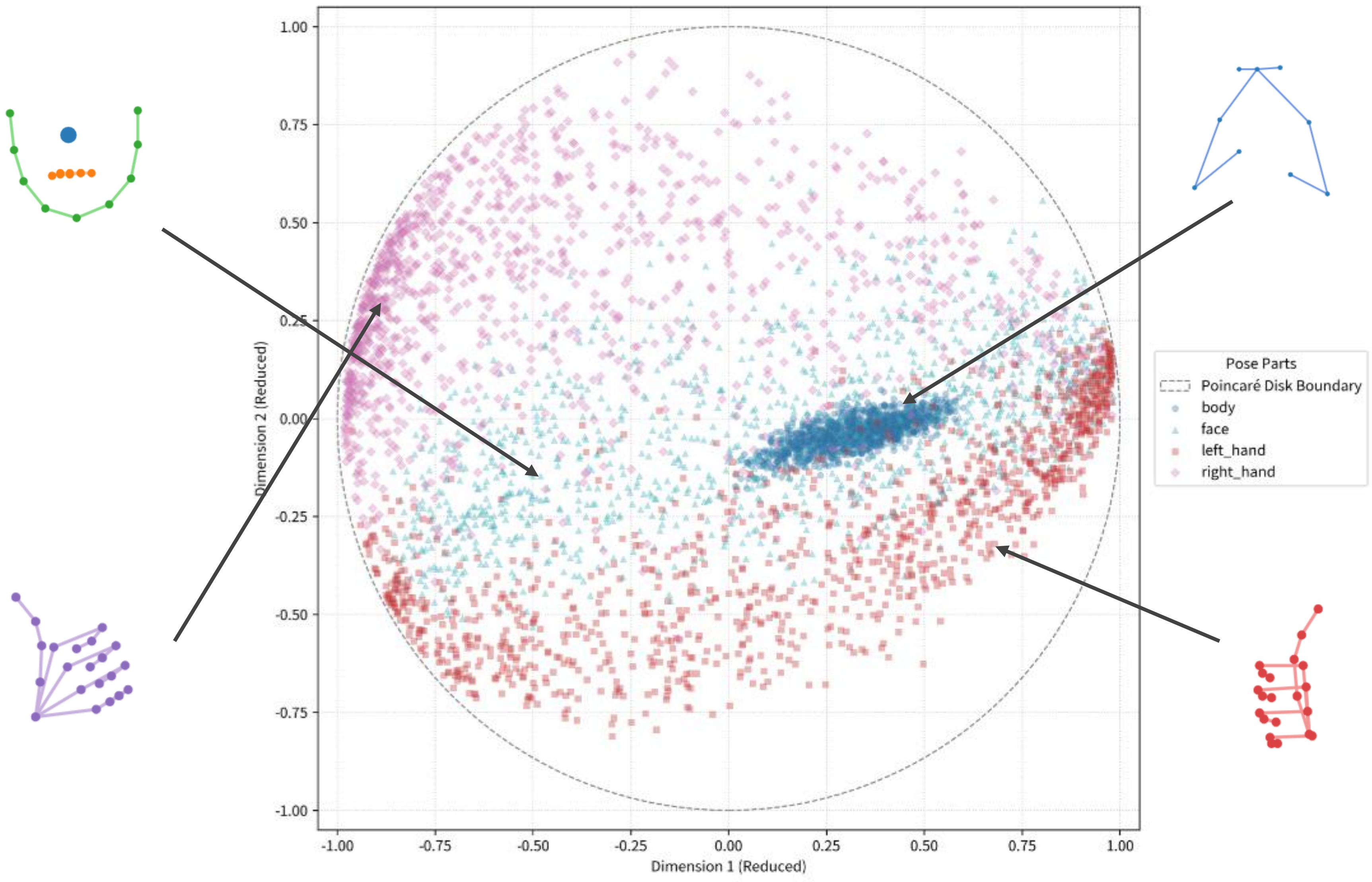
$$\mathcal{L}_{\text{hyp\_pair}}(\mathbf{p}_i, \mathbf{t}_i) = -\log \frac{\exp(-d_{\mathbb{B}_c}(\mathbf{p}_i, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{B} \exp(-d_{\mathbb{B}_c}(\mathbf{p}_i, \mathbf{t}_j)/\tau + m \cdot \mathbb{I}(i \neq j))}.$$

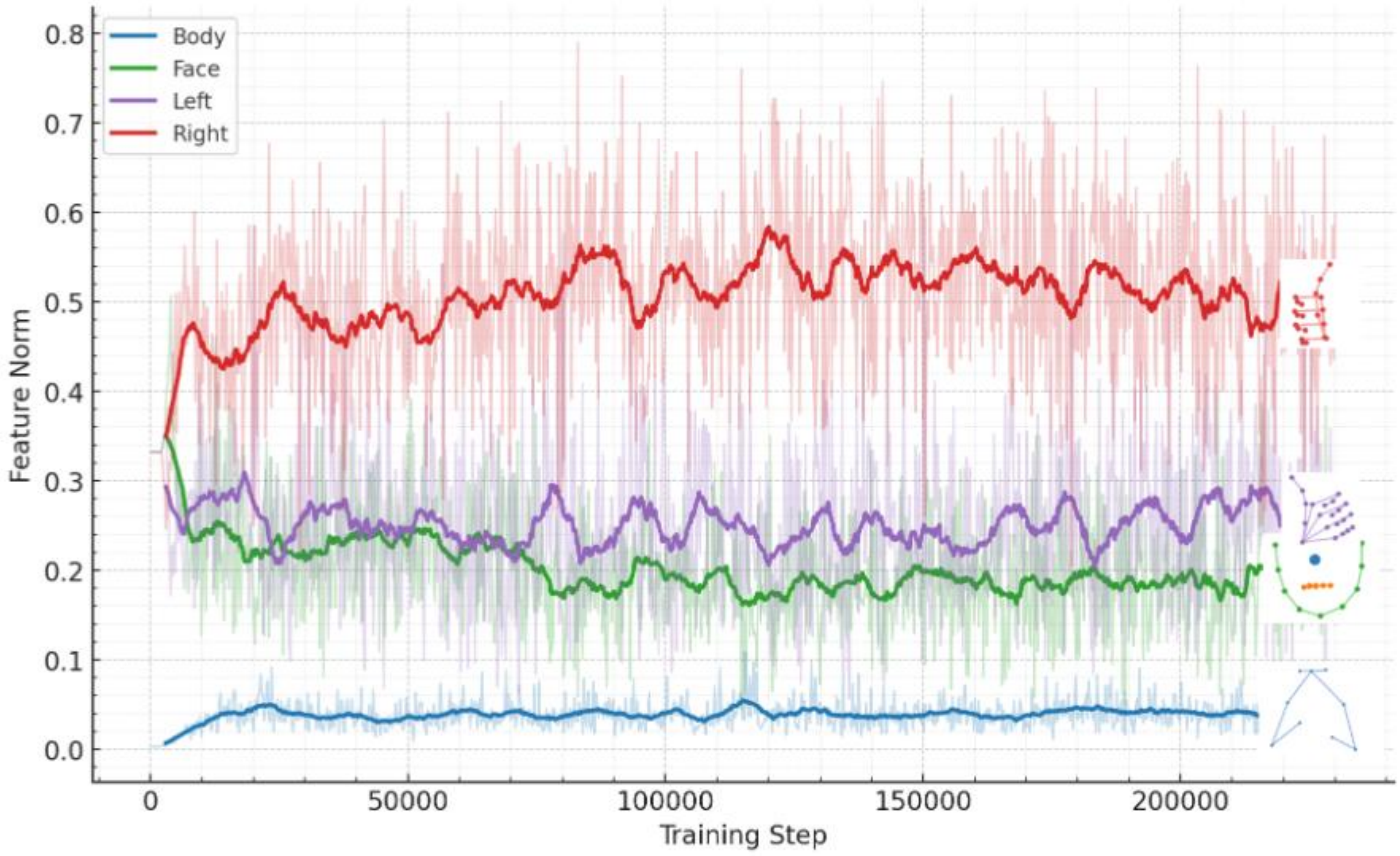




RH ST-GCN
LH ST-GCN
P ST-GCN
F ST-GCN
Pooled
exp map
Pool
Linear
Token
$c_{p_i} = \mu_{\mathbb{B}_c}(\{v_j\}_{j=1}^T, \{\alpha_{ij}\}_{j=1}^T)$
softmax
Hyp Attn
$(M \otimes_c v_i) \oplus_c b$
$\alpha_{i,j} = -d_{\mathbb{B}_c}(h_{p_i}, k_j)$
exp map
$-d(\mu, t)$
Token
Pooled
Translate:
Fusion
MHA
Add & Norm
FF
Add & Norm
Linear
Softmax
MT5
Translate: Today is the..


Pose Parts
Poincaré Disk Boundary
body
face
left_hand
right_hand
Dimension 1 (Reduced)
Dimension 2 (Reduced)
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00


Pose Parts
Poincaré Disk Boundary
body
face
left_hand
right_hand
Dimension 1 (Reduced)
Dimension 2 (Reduced)
-1.00
-0.75
-0.50
-0.25
0.00
0.25
0.50
0.75
1.00


Body
Face
Left
Right
Feature Norm
Training Step
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
50000
100000
150000
200000

Method	Modality		Dev Set			Test Set		
	Pose	RGB	B-1	B-4	R-L	B-1	B-4	R-L
Gloss-Based Methods (Prior Art)								
SLRT [6]	–	✓	37.47	11.88	37.96	37.38	11.79	36.74
TS-SLT [11]	✓	✓	55.21	25.76	55.10	55.44	25.79	55.72
CV-SLT [85]	–	✓	–	28.24	56.36	58.29	28.94	57.06
Gloss-Free Methods (Prior Art)								
MSLU [90]	✓	–	33.28	10.27	33.13	33.97	11.42	33.80
SLRT [6] (Gloss-Free variant)	–	✓	21.03	4.04	20.51	20.00	3.03	19.67
GASLT [83]	–	✓	–	–	–	19.90	4.07	20.35
GFSLT-VLP [88]	–	✓	39.20	11.07	36.70	39.37	11.00	36.44
FLa-LLM [10]	–	✓	–	–	–	37.13	14.20	37.25
Sign2GPT [74]	–	✓	–	–	–	41.75	15.40	42.36
SignLLM [19]	–	✓	42.45	12.23	39.18	39.55	15.75	39.91
C^2RL [9]	–	✓	–	–	–	49.32	21.61	48.21
Our Models and Baselines								
Uni-Sign [41] (Pose)	✓	–	53.24	25.27	54.34	53.86	25.61	54.92
Uni-Sign [41] (Pose+RGB)	✓	✓	55.30	26.25	56.03	55.08	26.36	56.51
Geo-Sign (Euclidean Pooled)	✓	–	53.53	25.78	55.38	53.06	25.72	55.57
Geo-Sign (Euclidean Token)	✓	–	53.93	25.91	55.20	54.02	25.98	53.93
Geo-Sign (Hyperbolic Pooled)	✓	–	55.19	26.90	56.93	55.80	27.17	57.75
Geo-Sign (Hyperbolic Token)	✓	–	**55.57**	**27.05**	**57.27**	**55.89**	**27.42**	**57.95**

Method	Modality		Dev Set			Test Set		
	Pose	RGB	B-1	B-4	R-L	B-1	B-4	R-L
Gloss-Based Methods (Prior Art)								
SLRT [6]	–	✓	37.47	11.88	37.96	37.38	11.79	36.74
TS-SLT [11]	✓	✓	55.21	25.76	55.10	55.44	25.79	55.72
CV-SLT [85]	–	✓	–	<u>28.24</u>	56.36	<u>58.29</u>	<u>28.94</u>	57.06
Gloss-Free Methods (Prior Art)								
MSLU [90]	✓	–	33.28	10.27	33.13	33.97	11.42	33.80
SLRT [6] (Gloss-Free variant)	–	✓	21.03	4.04	20.51	20.00	3.03	19.67
GASLT [83]	–	✓	–	–	–	19.90	4.07	20.35
GFSLT-VLP [88]	–	✓	39.20	11.07	36.70	39.37	11.00	36.44
FLa-LLM [10]	–	✓	–	–	–	37.13	14.20	37.25
Sign2GPT [74]	–	✓	–	–	–	41.75	15.40	42.36
SignLLM [19]	–	✓	42.45	12.23	39.18	39.55	15.75	39.91
C^2RL [9]	–	✓	–	–	–	49.32	21.61	48.21
Our Models and Baselines								
Uni-Sign [41] (Pose)	✓	–	53.24	25.27	54.34	53.86	25.61	54.92
Uni-Sign [41] (Pose+RGB)	✓	✓	55.30	26.25	56.03	55.08	26.36	56.51
Geo-Sign (Euclidean Pooled)	✓	–	53.53	25.78	55.38	53.06	25.72	55.57
Geo-Sign (Euclidean Token)	✓	–	53.93	25.91	55.20	54.02	25.98	53.93
Geo-Sign (Hyperbolic Pooled)	✓	–	55.19	26.90	56.93	55.80	27.17	57.75
Geo-Sign (Hyperbolic Token)	✓	–	**55.57**	**27.05**	**57.27**	**55.89**	**27.42**	**57.95**

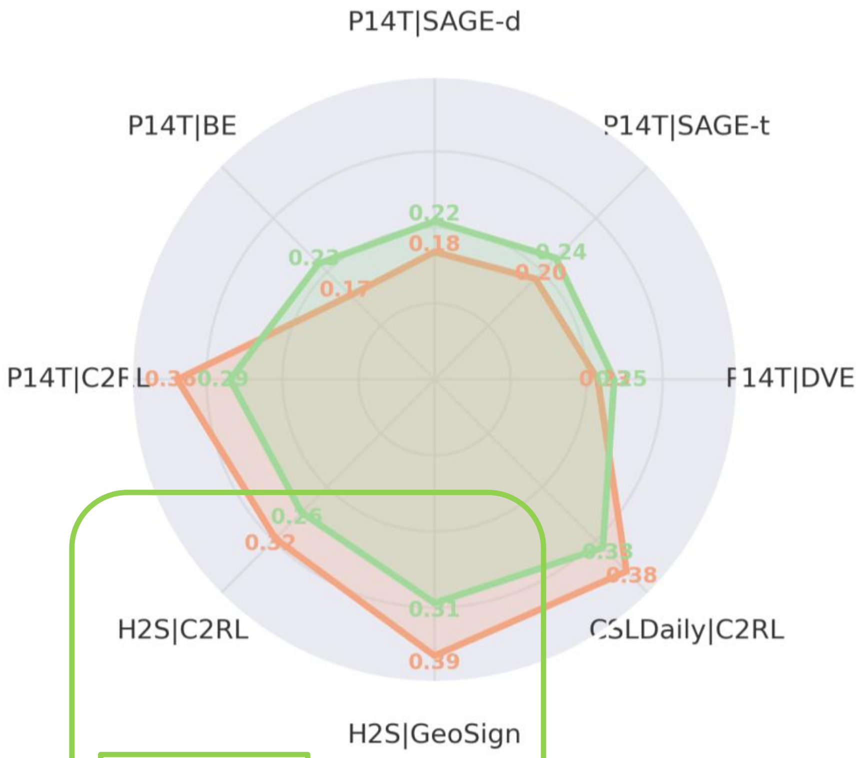
+3.03

Table 10: Comparative Analysis: Geo-Sign (Token) vs. Uni-Sign (Pose) - Selected Examples

Reference (Ground Truth)	Geo-Sign (Token) Prediction	Uni-Sign (Pose) Prediction
‘他每天回来都很累。‘ (He is very tired every day when he comes back.)	‘他每天来很累。‘ (He comes very tired every day.)	‘他每天来得及很累。‘ (He has enough time [to be/and is] very tired every day.)
‘小张,那个女生是你们公司的吗?你对她了解吗¿ (Xiao Zhang, is that girl from your company? Do you know her?)	‘小张那个女生是你公司的吗¿ (Xiao Zhang, is that girl from your company?)	‘那个小张是这家公司负责人,你了解吗¿ (That Xiao Zhang is the person in charge of this company, do you understand/know?)
‘阴天,电视上说多云,怎么了?明天有事¿ (Cloudy day, TV says it’s overcast, what’s up? Got plans tomorrow?)	‘阴天说什么话?天气什么的,明天有事。‘ (What to say on a cloudy day? Weather something, have things to do tomorrow.)	‘阴阳怪气地讲着天赋,不知不觉就发生了什么。‘ (Sarcastically talking about talent, something happened unknowingly.)
‘但是你一点也没瘦,你做什么运动了¿ (But you haven’t lost any weight, what exercise have you been doing?)	‘但是你没有太吃饱,你去做什么运动¿ (But you didn’t eat too full, what exercise are you going to do?)	‘但是你已经吃不消了,你能做什么呢¿ (But you already can’t stand it, what can you do?)

Continued on next page

All Verbs — Exact Match (%)



+.7

(a) Impact of Curvature c.

Curvature (c)	B-4	R-L
0.00 (Euclidean)	25.98	53.93
0.10	26.56	57.56
0.50	26.34	56.30
1.00	27.04	57.67
1.50	**27.42**	57.95
2.00	27.25	**58.08**

(b) Impact of α.

α	B-4	R-L
0.10	25.74	56.20
0.50	26.79	57.38
0.70	**27.42**	**57.95**
0.90	26.92	57.67

(c) Noise Robustness (B-4).

Noise	Geo-Sign	Uni-Sign
0.00	27.42	26.25
0.01	26.30 (-4%)	24.14 (-8%)
0.02	24.60 (-10%)	21.50 (-18%)
0.03	19.07 (-30%)	14.40 (-45%)
0.04	11.63 (-58%)	7.20 (-73%)
0.05	5.98 (-78%)	3.01 (-89%)

github.com/ed-fish/geo-sign
edward.fish@surrey.ac.uk

Geo-Sign: Hyperbolic Contrastive Regularisation for Geometrically Aware Sign Language Translation

Edward Fish
CVSSP, University of Surrey
edward.fish@surrey.ac.uk

Richard Bowden
CVSSP, University of Surrey
r.bowden@surrey.ac.uk

arXiv:2506.00129v2 [cs.CV] 28 Oct 2025

Abstract

Recent progress in Sign Language Translation (SLT) has focussed primarily on improving the representational capacity of large language models to incorporate Sign Language features. This work explores an alternative direction: enhancing the geometric properties of skeletal representations themselves. We propose Geo-Sign, a method that leverages the properties of hyperbolic geometry to model the hierarchical structure inherent in sign language kinematics. By projecting skeletal features derived from Spatio-Temporal Graph Convolutional Networks (ST-GCNs) into the Poincaré ball model, we aim to create more discriminative embeddings, particularly for fine-grained motions like finger articulations. We introduce a hyperbolic projection layer, a weighted Fréchet mean aggregation scheme, and a geometric contrastive loss operating directly in hyperbolic space. These components are integrated into an end-to-end translation framework as a regularisation function, to enhance the representations within the language model. This work demonstrates the potential of hyperbolic geometry to improve skeletal representations for Sign Language Translation, improving on SOTA RGB methods while preserving privacy and improving computational efficiency. Code available here: https://github.com/ed-fish/geo-sign.

1 Introduction

Sign Languages are rich, multi-channel linguistic systems where meaning is conveyed through a composition of movements involving the upper body, hands, face, and mouth. Automatic Sign Language Translation (SLT) is an established research area focused on developing methods to convert these visual expressions directly into text. While Sign Languages are expressed via fluid multi-articulator kinematics, a persistent challenge for SLT methods lies in creating feature representations that concurrently preserve fine-grained, local details (e.g., subtle finger configurations) while embedding the global structure inherent in larger, overarching body motions. Effectively modelling these multi-scale and relational dynamics within a suitable geometric embedding space remains a central hurdle.

Spatio-Temporal Graph Convolutional Networks (ST-GCNs) offer a natural way to encode these hierarchical relationships by treating the body's joints and bones as nodes and edges in a graph [79]. However, when their learned representations are projected into standard Euclidean geometry for processing via a Large Language Model (LLM), essential fine-grained relational distances and movements can become blurred. For instance, the sign for "*water*" in American Sign Language (ASL) is communicated by forming a W shape with the fingers and tapping the chin twice (a fine-grained, "leaf-level" articulation), immediately followed by a sweeping hand movement away from the body (a "branch-level" gesture). When these features are aggregated in Euclidean geometry, the large translation and rotation of the wrist could dominate the vector's norm, effectively "pulling" the embedding toward the global motion and compressing the subtle finger tap into a vanishing tail. Consequently, two signs that differ only in the timing or precision of that tap, which may be critical to lexical meaning, can become nearly indistinguishable once projected into flat Euclidean space.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The **British Machine Vision Association**
and **Society for Pattern Recognition**

One Day Meeting: BMVA Symposium on AI for Sign Language Translation, Production, and Linguistics

Wednesday 10 December 2025

Chairs: Edward Fish (University of Surrey), Youngjoon Jang (KAIST), Özge Mercanoğlu Sincan (University of Surrey)

Prof Andrew Zisserman

Prof Richard Bowden

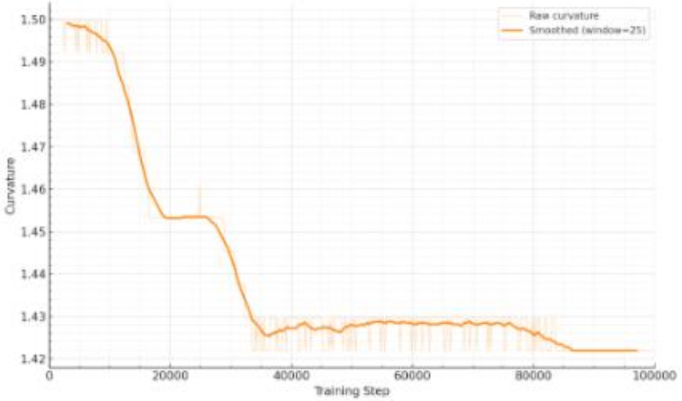
Dr Abraham Glasser

Algorithm 1 Iterative Weighted Fréchet Mean in $\mathbb{B}_c^{d_{\text{hyp}}}$

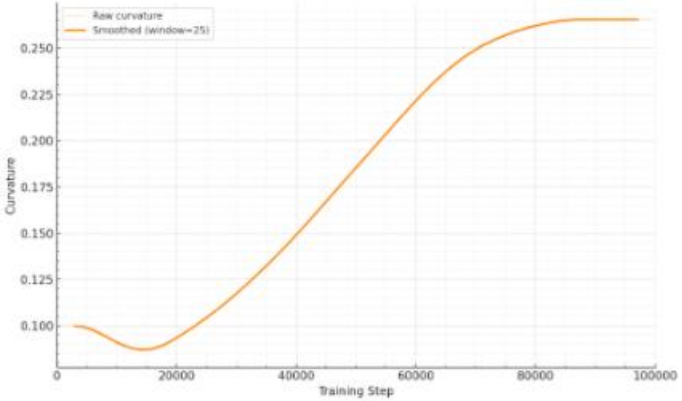
Require: Hyperbolic embeddings $\{\mathbf{h}_p\}_{p=1}^N$, normalized positive weights $\{w_p\}_{p=1}^N$, c, $I_{\max}$, ϵ_{tol}.

Initialize $\boldsymbol{\mu}^{(0)} \leftarrow \mathbf{h}_1$ (or other suitable initialization).
for $k = 0$ to $I_{\max} - 1$ **do**
  $\mathbf{v}_{\text{agg}} \leftarrow \mathbf{0} \in \mathcal{T}_{\boldsymbol{\mu}^{(k)}}\mathbb{B}_c^{d_{\text{hyp}}}$. ▷ Aggregated tangent vector at current mean
  for $p = 1$ to N **do**
    $\mathbf{v}_{\text{agg}} \leftarrow \mathbf{v}_{\text{agg}} + w_p \log^c_{\boldsymbol{\mu}^{(k)}}(\mathbf{h}_p)$. ▷ Sum weighted log-mapped vectors
  end for
  $\boldsymbol{\mu}^{(k+1)} \leftarrow \exp^c_{\boldsymbol{\mu}^{(k)}}(\mathbf{v}_{\text{agg}})$. ▷ Update mean via exponential map
  Project $\boldsymbol{\mu}^{(k+1)}$ into $\mathbb{B}_c^{d_{\text{hyp}}}$ if numerically necessary.
  if $d_{\mathbb{B}_c}(\boldsymbol{\mu}^{(k+1)}, \boldsymbol{\mu}^{(k)}) < \epsilon_{\text{tol}}$ **then** ▷ Check convergence
    break ▷ Exit loop on convergence
  end if
end for
$\boldsymbol{\mu}_{\text{pose}} \leftarrow \boldsymbol{\mu}^{(k+1)}$ ▷ Assign final mean

Ensure: Estimated Fréchet mean $\boldsymbol{\mu}_{\text{pose}}$.



(a) c initialised at 1.50, decreases and stabilizes around 1.42.



(b) c initialised at 0.10, increases and stabilizes around 0.20.

Figure 3: Evolution of the learnable manifold curvature c during training for different initializations. (a) When initialised at $c = 1.50$, the curvature magnitude slightly decreases, suggesting an optimal value around 1.42 for this setup. (b) When initialised at a low $c = 0.10$, the curvature increases, indicating the model benefits from more "hyperbolic space" initially. It stabilizes around $c = 0.20$, potentially influenced by the dynamic α schedule that reduces regularization emphasis over time.

Model Variant (Batch Size 8)	Total Params (M)	Added Params (M)	Total Fwd MACs (GMACs)	Hyperbolic Proj. Layer MACs (MMACs)	Fwd Latency (ms)	Latency Increase (%)
Baseline Uni-Sign (Pose)	587.75	-	116.59	-	415.73	-
Geo-Sign (Hyperbolic Pooled)	588.21	0.46	116.60	3.67	1630.00	292.10
Geo-Sign (Hyperbolic Token)	589.10	1.35	116.60	$\approx$9.96	2550.00	513.40

Method	VE Name	VE Params (M)	LM Name	LM Params (M)	Total Params (M)	Modality		Test Set	
						Pose	RGB	B-4	R-L
Gloss-Free Methods (Prior Art)									
MSLU [90]	EffNet	5.3	mT5-Base	582.4	587.7	✓	–	11.42	33.80
SLRT [6] (G-Free)	EffNet	5.3	Transformer	≈30	≈35.3	–	✓	3.03	19.67
GASLT [83]	I3D	13	Transformer	≈30	≈43.0	–	✓	4.07	20.35
GFSLT-VLP [88]	ResNet18	11.7	mBart	680	691.7	–	✓	11.00	36.44
FLa-LLM [10]	ResNet18	11.7	mBart	680	691.7	–	✓	14.20	37.25
Sign2GPT [74]	DinoV2	21.0	XGLM	1732.9	1753.9	–	✓	15.40	42.36
SignLLM [19]	ResNet18	11.7	LLaMA-7B	6738.4	6750.1	–	✓	15.75	39.91
C^2RL [9]	ResNet18	11.7	mBart	680	691.7	–	✓	21.61	48.21
Our Models and Baselines									
Uni-Sign [41] (Pose)	GCN	5.3	mT5-Base	582.4	587.7	✓	–	25.61	54.92
Uni-Sign [41] (Pose+RGB)	EffNet+GCN	9.7	mT5-Base	582.4	592.1	✓	✓	26.36	56.51
Geo-Sign (Hyperbolic Pooled)	GCN+Geo	5.8	mT5-Base	582.4	588.21	✓	–	27.17	57.75
Geo-Sign (Hyperbolic Token)	GCN+Geo+Attn	6.7	mT5-Base	582.4	589.1	✓	–	**27.42**	**57.95**

Category	Hyperparameter	Value	Description
General Training Configuration			
	Random Seed	42	Seed for reproducibility
	Training Epochs	40	Number of fine-tuning epochs on CSL-Daily
	Batch Size (per GPU)	8	Micro-batch size per GPU
	Gradient Accumulation Steps	8	Effective batch size becomes 8 × accum_steps × num_gpus
	Training Precision (`dtype`)	`bf16`	Mixed precision training data type
Data Handling			
	Max Pose Sequence Length	256	Maximum number of frames for pose sequences
	Max Target Text Length (`max_tgt_len`)	100	Max new tokens for generation during evaluation
Optimizer (Euclidean: ST-GCN, mT5, Linear Layers)			
	Optimizer Type (`opt`)	AdamW	[45]
	Learning Rate (`lr`)	3×10^{-5}	For Euclidean parameters (AdamW)
	AdamW β_1, β_2 (`opt-betas`)	[0.9, 0.999]	Exponential decay rates for moment estimates
	AdamW ϵ (`opt-eps`)	1×10^{-8}	Term for numerical stability
	Weight Decay (`weight-decay`)	0.01	L2 penalty for Euclidean parameters
	LR Scheduler (`sched`)	Cosine Annealing	
	Warmup Epochs (`warmup-epochs`)	5	Number of epochs for LR warm-up
	Minimum LR (`min-lr`)	1×10^{-6}	Lower bound for LR in scheduler
	Gradient Clipping Norm	1.0	Max norm for gradients
Optimizer (Hyperbolic: Manifold Parameters, Projections)			
	Optimizer Type	RAdam	Riemannian Adam
	Learning Rate (`hyp_lr`)	1×10^{-3}	For hyperbolic parameters (RAdam)
Model Architecture			
	ST-GCN Output Dimension (`gcn_out_dim`)	256	Output dimension of ST-GCN part streams
	mT5 Projection Dimension (`hidden_dim`)	768	Target dimension for projecting GCN features to match mT5
Hyperbolic Regularization			
	Hyperbolic Embedding Dimension (d_{hyp}, `hyp_dim`)	256	Dimension of embeddings in Poincaré ball
	Initial Curvature (c_{init}, `init_c`)	1.5	Initial value for learnable curvature c (for best model)
	Loss Blend α_{init} (`alpha`)	0.70	Initial blending factor for $\mathcal{L}_{\text{CE}}$ vs $\mathcal{L}_{\text{hyp\_reg}}$ (for best model)
	Text Comparison Mode (`hyp_text_cmp`)	`token`	Strategy for aligning pose with text tokens (Token Method)
	Hyperbolic Contrastive Loss $\mathcal{L}_{\text{hyp\_reg}}$:		
	Temperature (τ)	Learnable	Temperature for scaling distances in contrastive loss
	Margin (m)	Learnable	Additive margin for negative pairs in contrastive loss
	Label Smoothing (`label_smoothing_hyp`)	0.2	Label smoothing for hyperbolic contrastive loss (InfoNCE)
Loss Functions			
	CE Loss Label Smoothing (`label_smoothing`)	0.2	Label smoothing for mT5 cross-entropy loss
Distributed Training (DeepSpeed)			
	ZeRO Optimization Stage (`zero_stage`)	2	DeepSpeed ZeRO Stage for memory efficiency
	Offload to CPU (`offload`)	False	Whether to offload optimizer/params to CPU