ICCV HONOLULU HAWAII
OCT 19-23, 2025

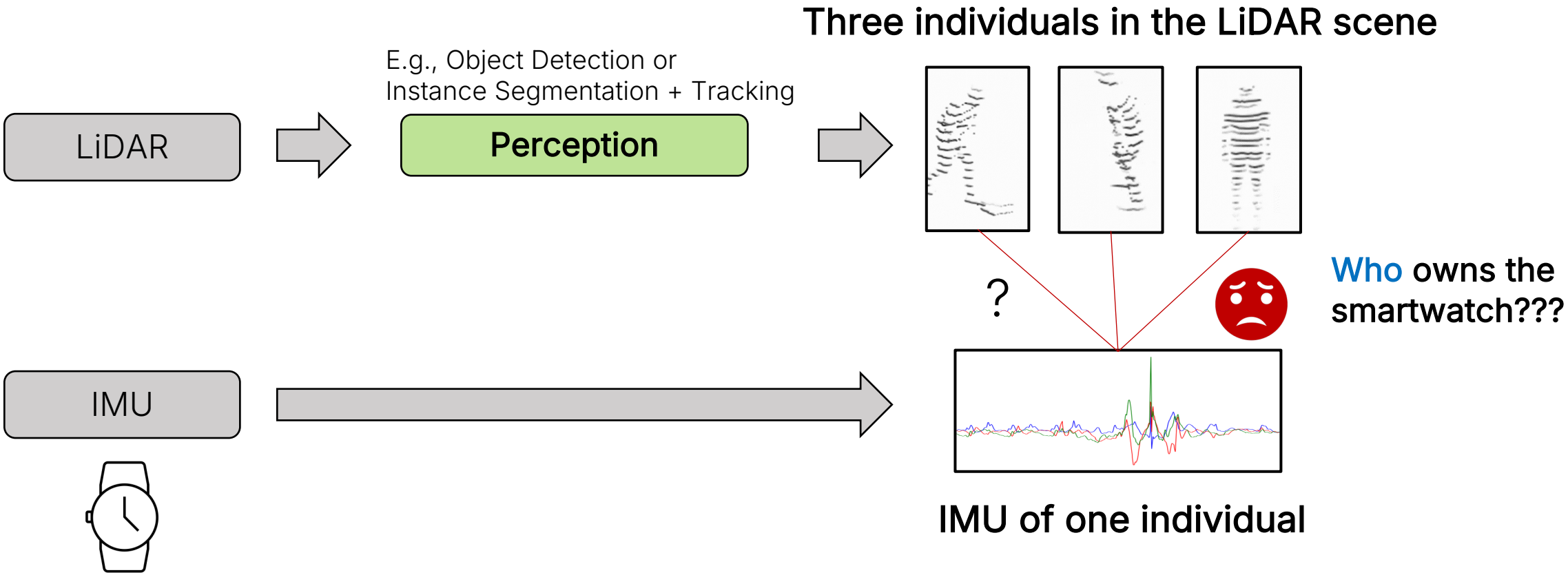
TECHNISCHE UNIVERSITÄT DARMSTADT

DeSPITE: Exploring Contrastive Deep Skeleton-Pointcloud-IMU-Text Embeddings for Advanced Point Cloud Human Activity Understanding

Thomas Kreutz, Max Mühlhäuser, Alejandro Sanchez Guinea
TU Darmstadt, Telecooperation Lab

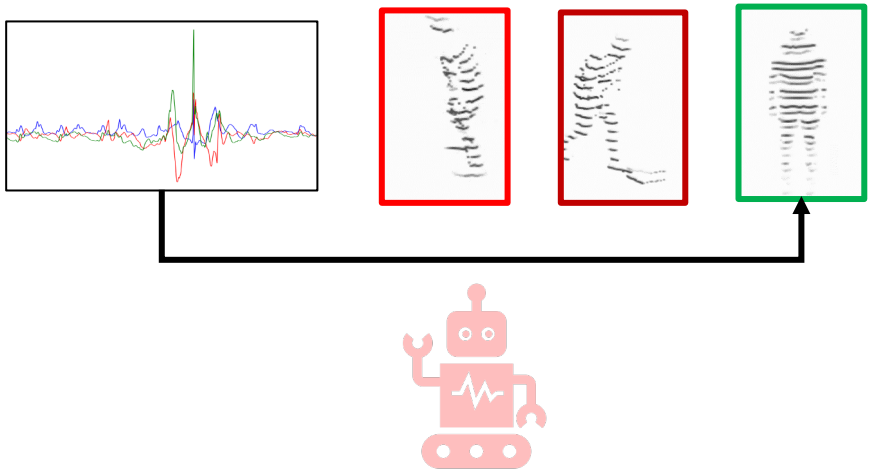
Telecooperation Lab

The problem: In the real world LiDAR and IMU are independent of each other

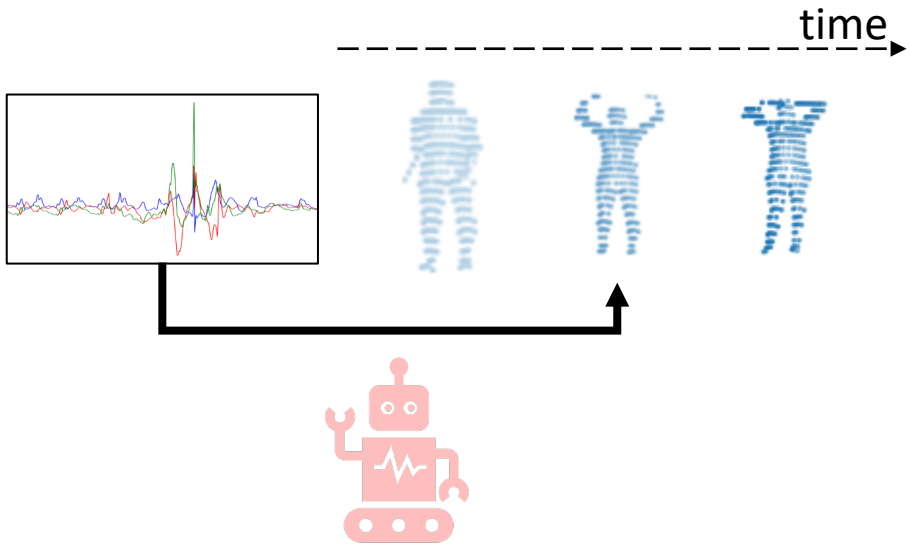


Our Goal:
Learn how LiDAR data corresponds to IMU data

Who is this?

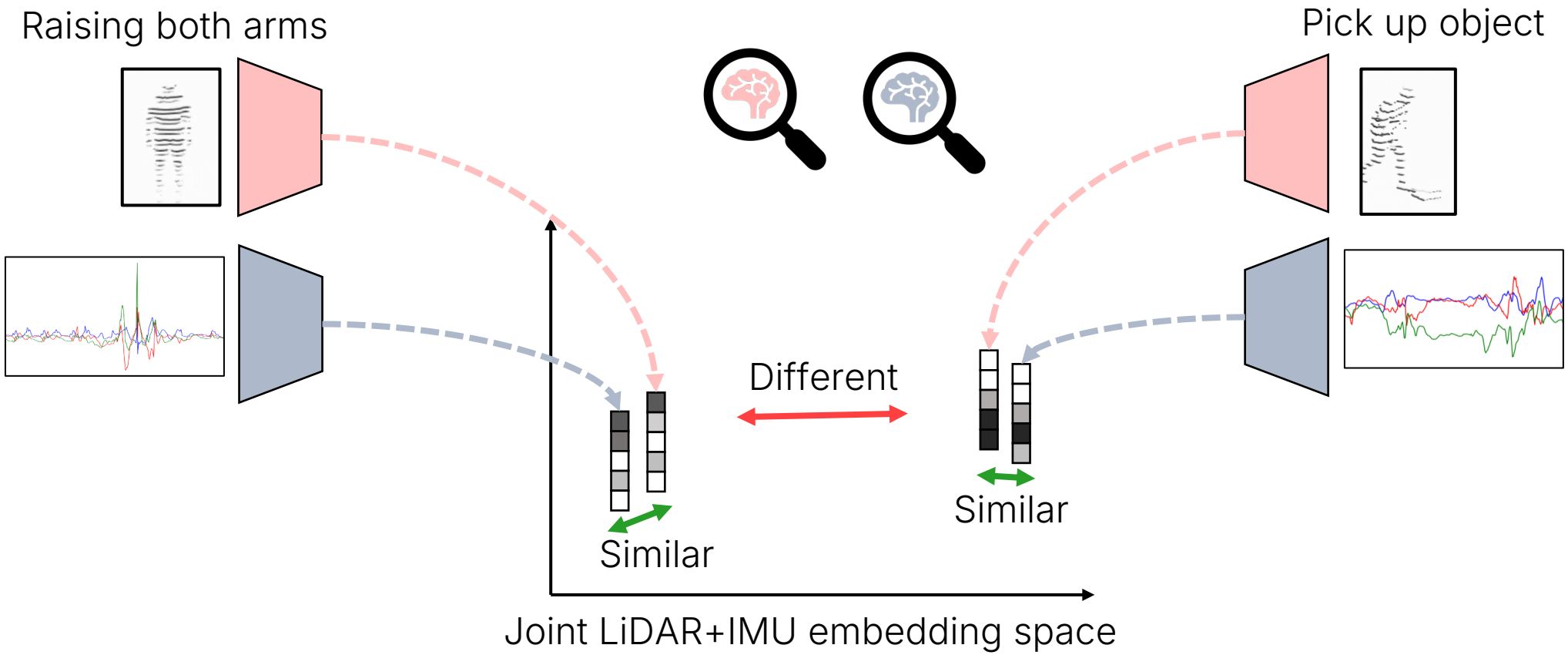


When did this happen?



Research Question

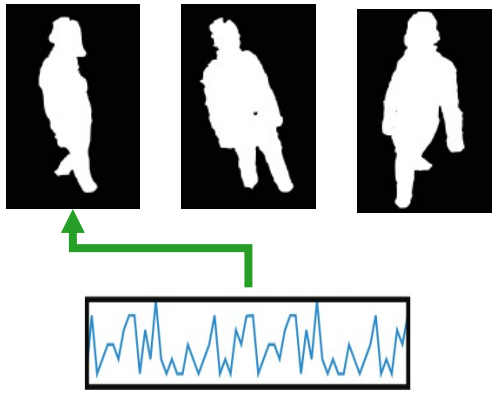
Can we use multi-modal representation learning to learn the correspondence between LiDAR and IMU data?



Related Work: Multi-Modal Representation Learning

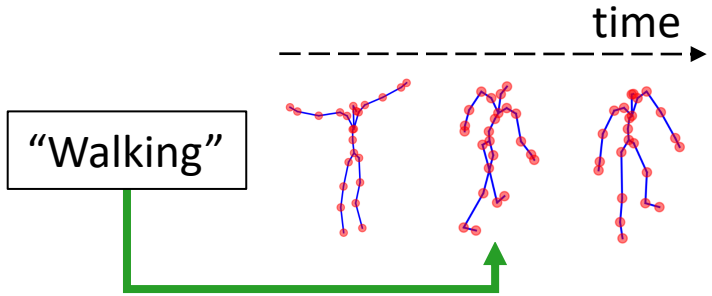
- **Multi-modal representation learning is great for correspondence problems**

IMU+RGB for RE-ID[1]

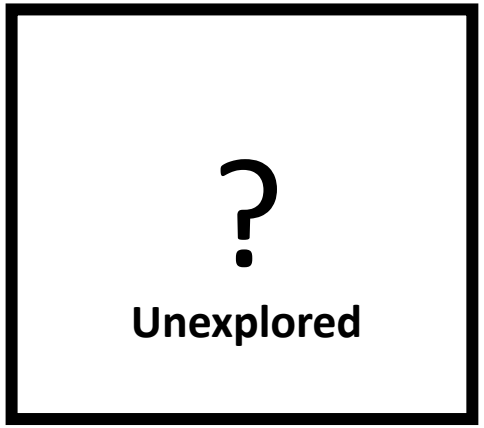


IMU+RGB
for search with IMU and Interpretability[2]

Text+Skeleton pose
for search with text query[3]



IMU+LiDAR?



1 Masullo et al., 2019, Who Goes There? Exploiting Silhouettes and Wearable Signals for Subject Identification in Multi-Person Environments
2 Moon et al., 2022, IMU2CLIP: Multimodal Contrastive Learning for IMU Motion Sensors from Egocentric Videos and Text
3 Petrovich et al., 2023, TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis

Related Work: Multi-Modal Representation Learning

- **Multi-modal representation learning is great for correspondence problems**
- **Joint embedding spaces benefit from more modalities (even if only partial alignments between modalities)**

- No LiDAR

ImageBind[1]

- RE-ID and search not explored

BabelTower[2]

1 Girdhar et al., 2023, ImageBind: One Embedding Space To Bind Them All
2 Dai et al., 2024, Babel: A Scalable Pre-trained Model for MultiModal Sensing via Expandable Modality Alignment

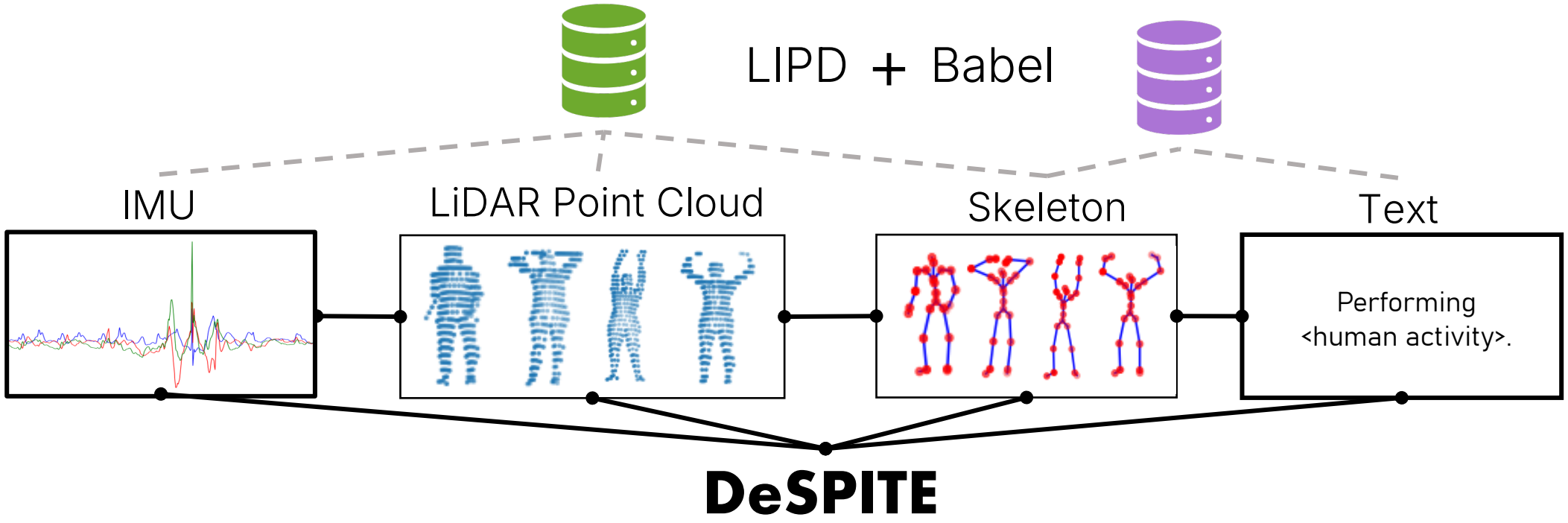
Contributions

Dataset: LIPD[1]+Babel[2]

- ~500.000 synchronized Point Cloud, IMU, Skeleton with partial text descriptions

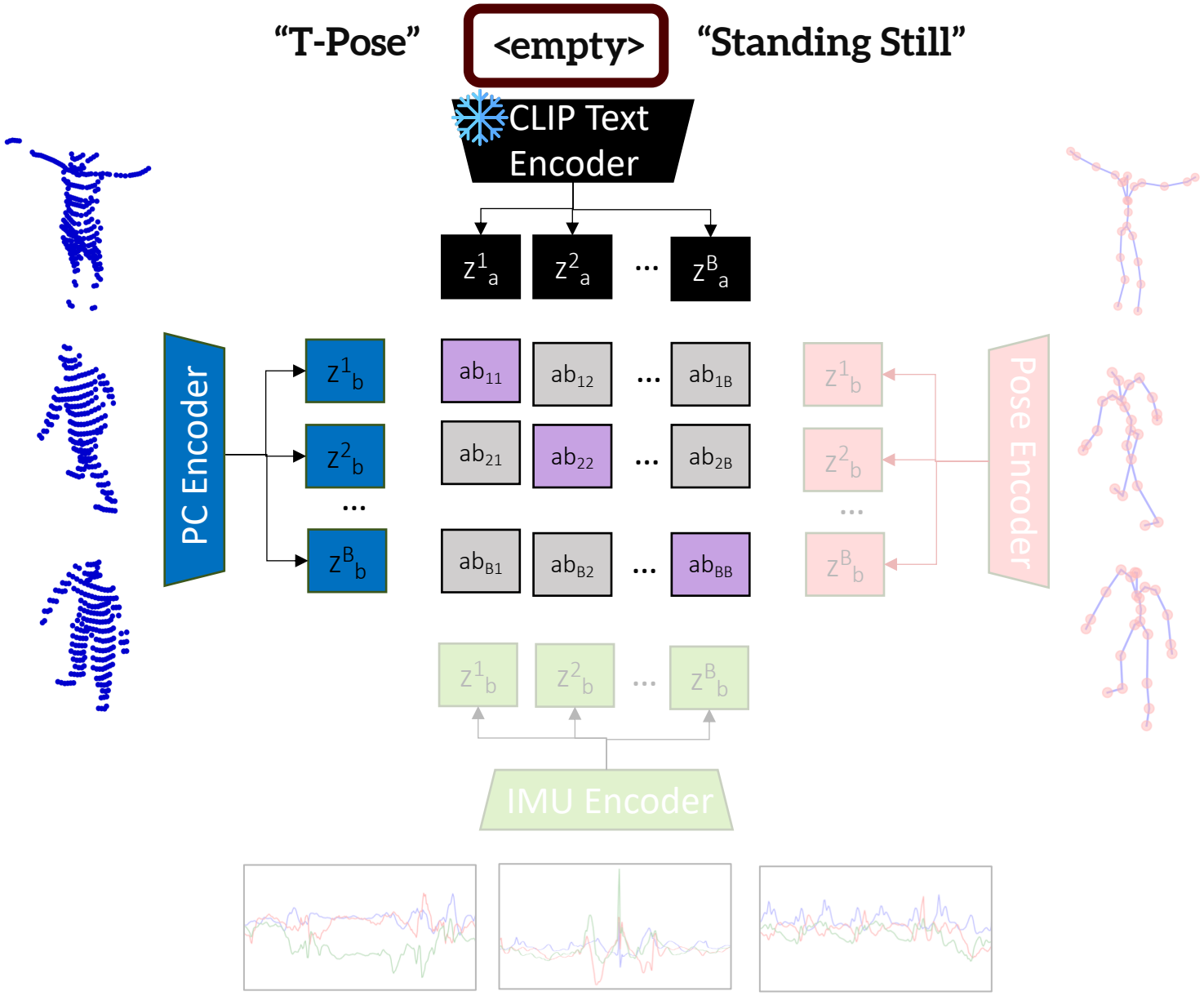
DeSPITE = **De**ep **S**keleton **P**ointcloud **I**MU **T**ext **E**mbeddings

- Joint embedding space through multi-modal contrastive learning



1 Ren et al., 2022, LiDAR-aid Inertial Poser: Large-scale Human Motion Capture by Sparse Inertial and LiDAR Sensors
2 Punnakkal et al., 2021, BABEL: Bodies, Action and Behavior with English Labels

Learning Joint Embedding Space



1. Bind to CLIP Text Embedding Space

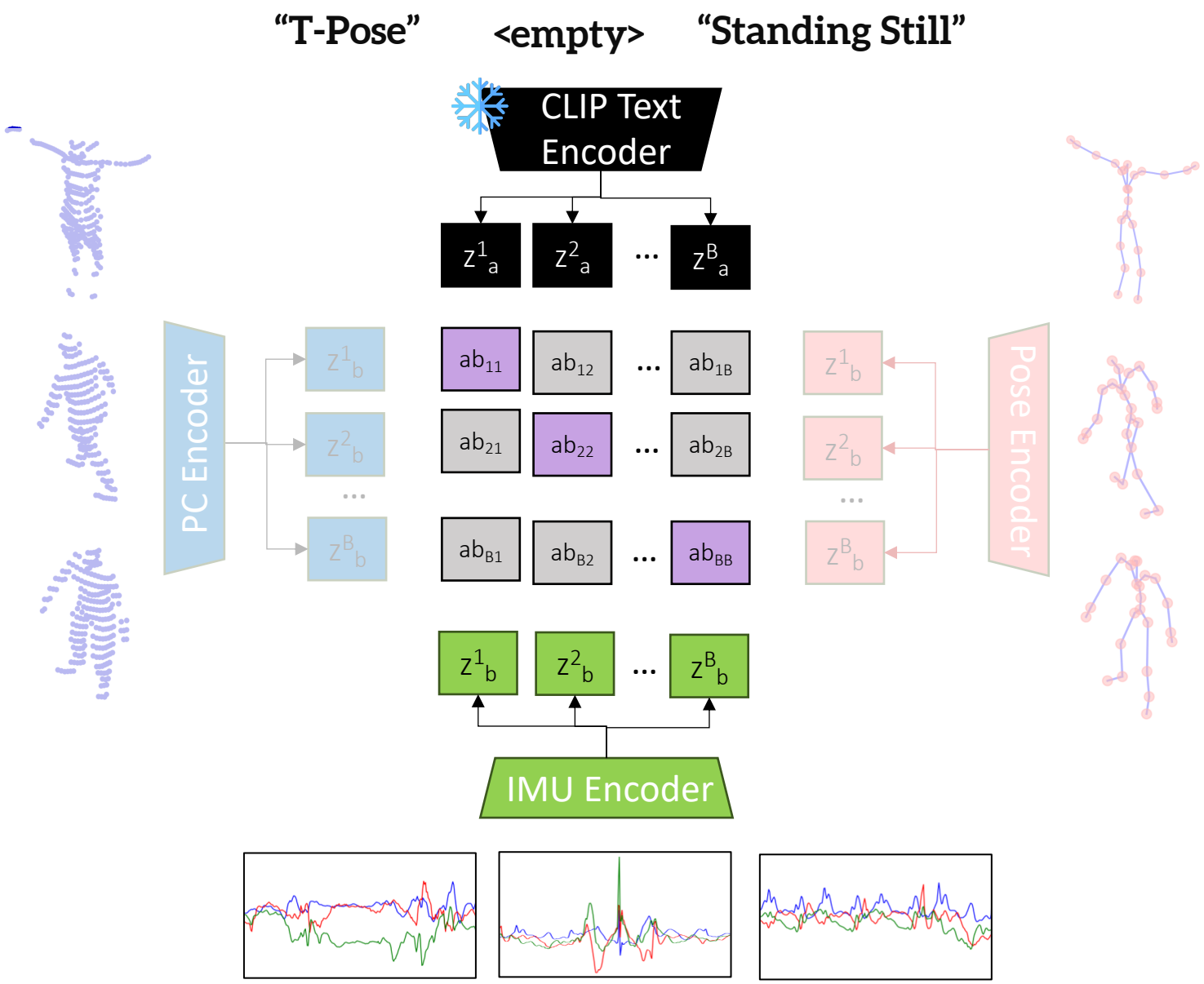
$$\mathcal{L}^i_{a\to b} = -\log \frac{\exp(\mathrm{sim}(z^i_a, z^i_b)/\tau)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(z^i_a, z^j_b)/\tau)}$$

Sum over all modality pairs!

$$\mathcal{L}_{text} = \sum_{i=1}^{B} \boxed{tm^i} \sum_{a\in M} \mathcal{L}^i_{a,text}$$

$\boxed{tm^i}$ mask if no text pairing exists (i.e., <empty>)

Learning Joint Embedding Space



1. Bind to CLIP Text Embedding Space

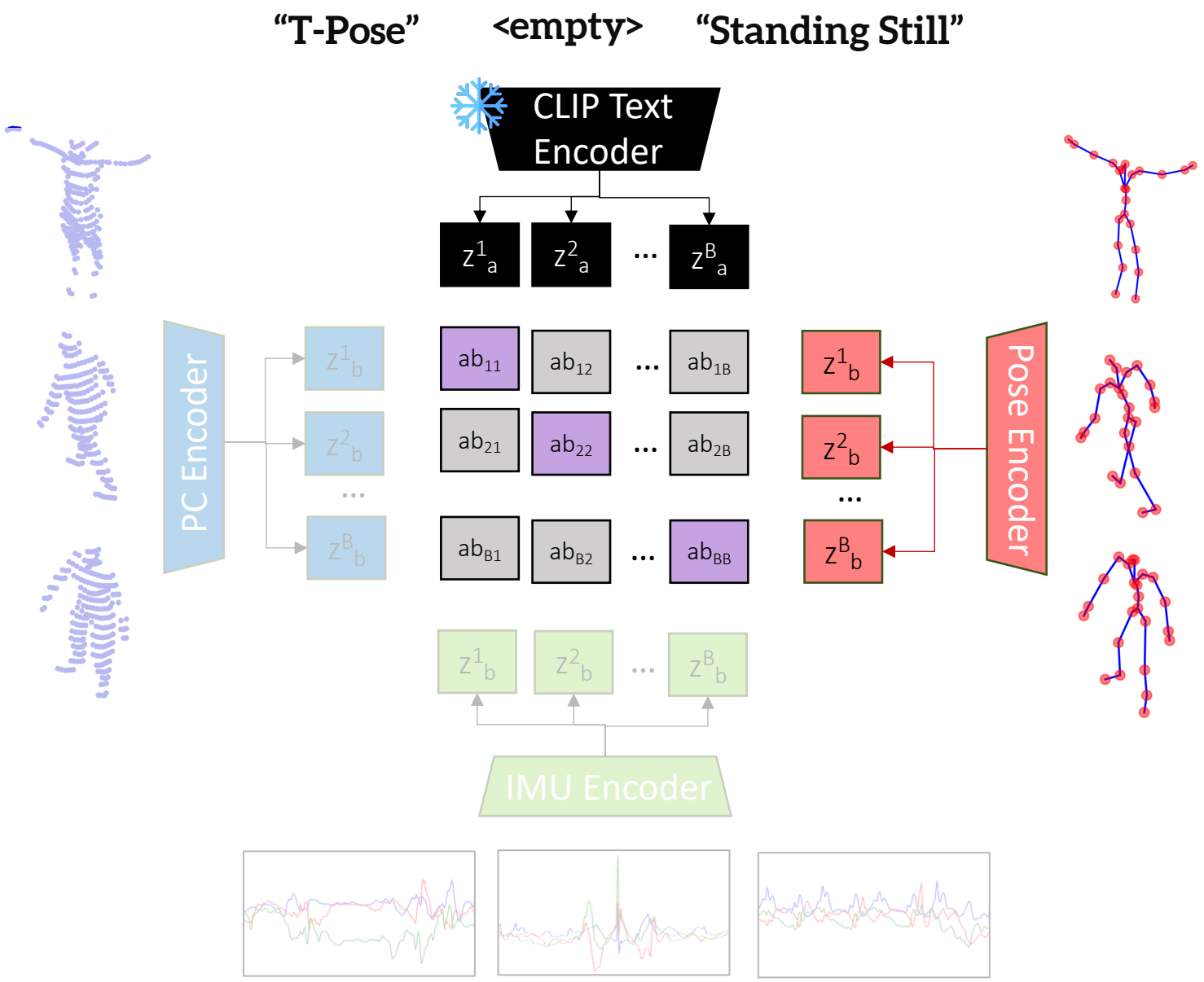
$$\mathcal{L}^i_{a\to b} = -\log \frac{\exp(\mathrm{sim}(z^i_a, z^i_b)/\tau)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(z^i_a, z^j_b)/\tau)}$$

Sum over all modality pairs!

$$\boxed{\mathcal{L}_{text} = \sum_{i=1}^{B} tm^i \sum_{a\in M} \mathcal{L}^i_{a,text}}$$

tm^i mask if no text pairing exists (i.e., <empty)

Learning Joint Embedding Space



1. Bind to CLIP Text Embedding Space

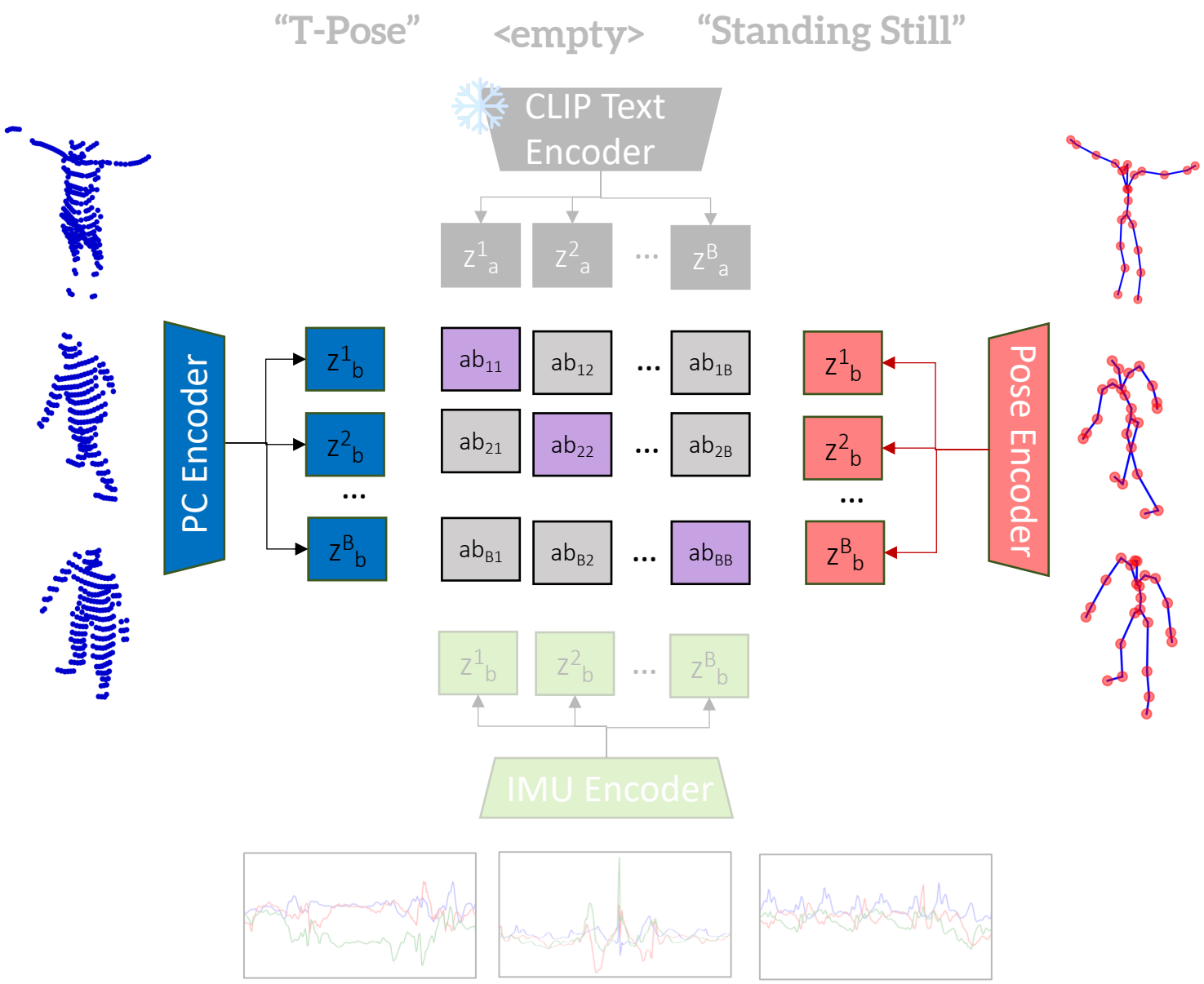
$$\mathcal{L}^i_{a\to b} = -\log \frac{\exp(\mathrm{sim}(z^i_a, z^i_b)/\tau)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(z^i_a, z^j_b)/\tau)}$$

Sum over all modality pairs!

$$\boxed{\mathcal{L}_{text} = \sum_{i=1}^{B} tm^i \sum_{a\in M} \mathcal{L}^i_{a,text}}$$

tm^i mask if no text pairing exists (i.e., <empty)

Learning Joint Embedding Space



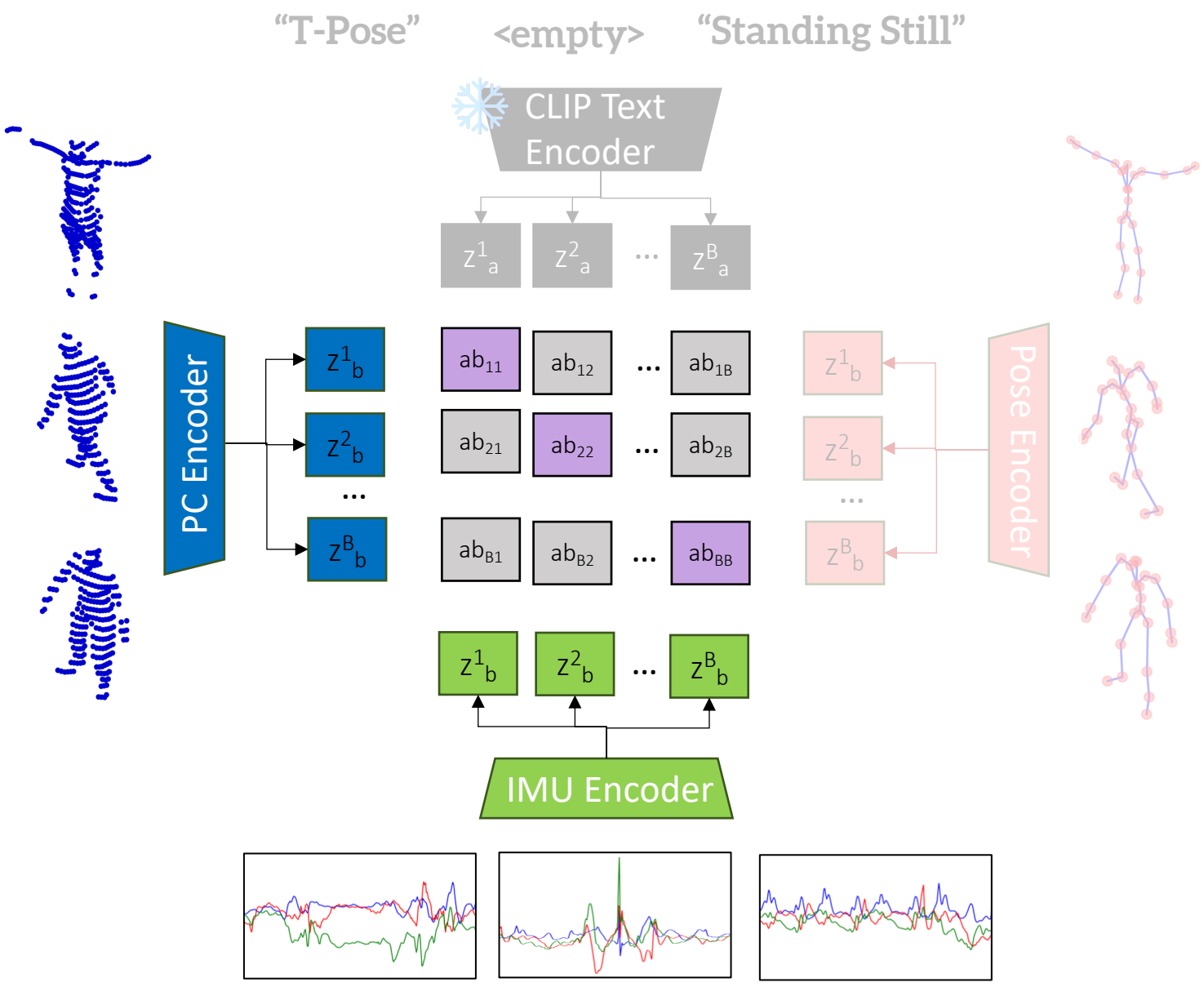
1. Bind to CLIP Text Embedding Space
2. **Pairwise non-text modality alignment**

$$\mathcal{L}^i_{a\to b} = -\log \frac{\exp(\mathrm{sim}(z^i_a, z^i_b)/\tau)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(z^i_a, z^j_b)/\tau)}$$

Sum over all modality pairs!

$$\boxed{\mathcal{L}_M = \sum_{i=1}^{B} \sum_{(a,b)\in M^*} \mathcal{L}^i_{a,b}}$$

Learning Joint Embedding Space



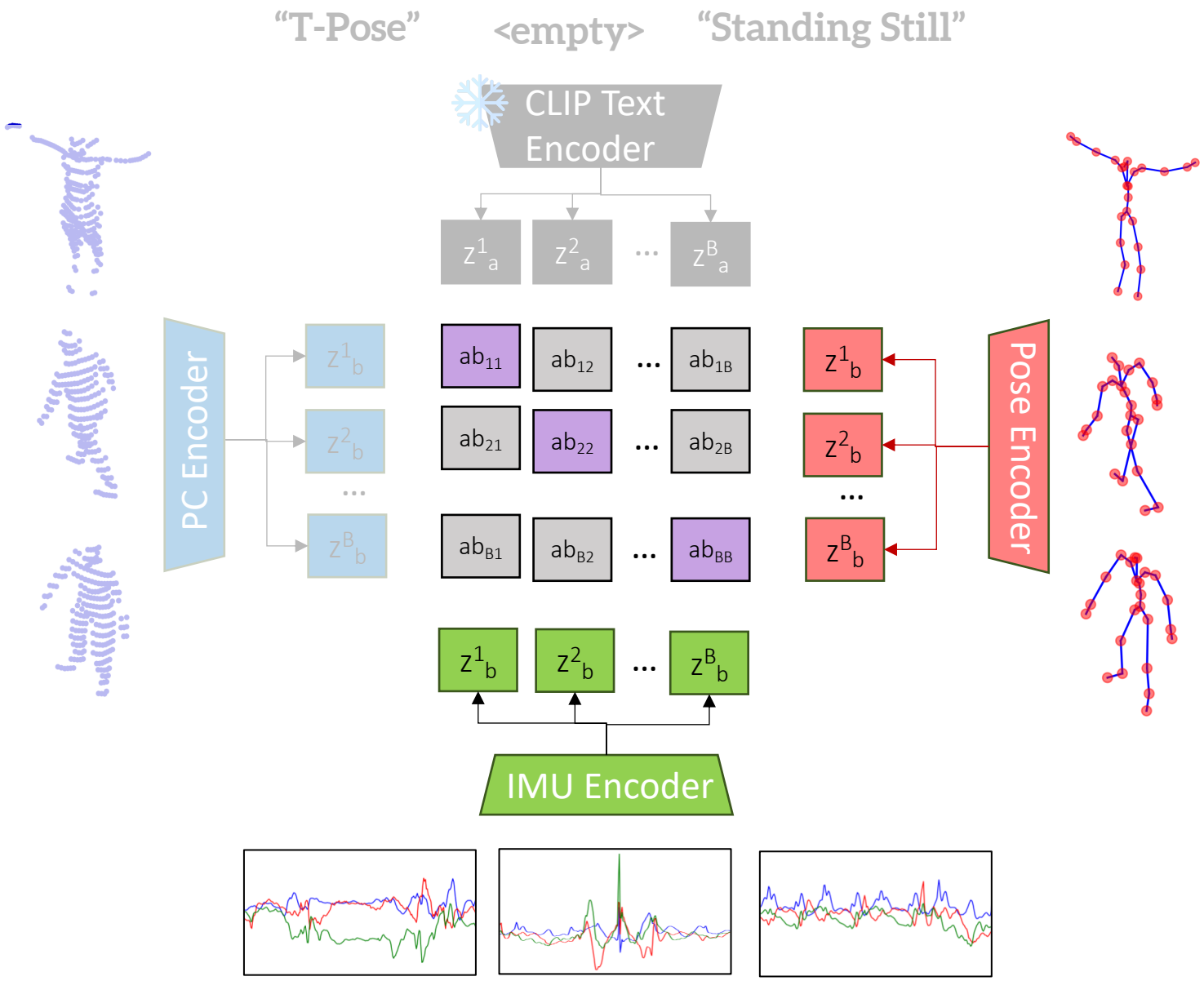
1. Bind to CLIP Text Embedding Space
2. **Pairwise non-text modality alignment**

$$\mathcal{L}^i_{a\to b} = -\log \frac{\exp(\mathrm{sim}(z^i_a, z^i_b)/\tau)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(z^i_a, z^j_b)/\tau)}$$

Sum over all modality pairs!

$$\boxed{\mathcal{L}_M = \sum_{i=1}^{B} \sum_{(a,b)\in M^*} \mathcal{L}^i_{a,b}}$$

Learning Joint Embedding Space



1. Bind to CLIP Text Embedding Space
2. **Pairwise non-text modality alignment**

$$\mathcal{L}^i_{a\to b} = -\log \frac{\exp(\mathrm{sim}(z^i_a, z^i_b)/\tau)}{\sum_{j=1}^{B} \exp(\mathrm{sim}(z^i_a, z^j_b)/\tau)}$$

Sum over all modality pairs!

$$\boxed{\mathcal{L}_M = \sum_{i=1}^{B} \sum_{(a,b)\in M^*} \mathcal{L}^i_{a,b}}$$

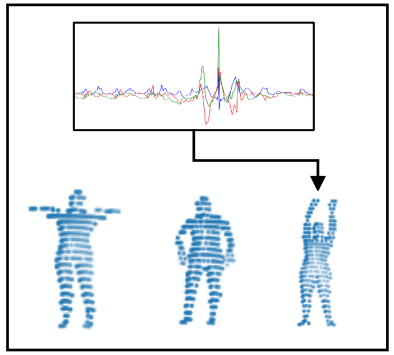
Evaluation

Evaluate on merged LIPD+Babel:

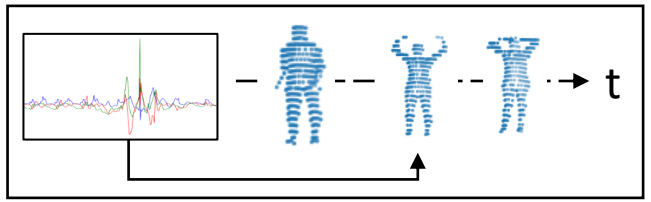
- 24 frame long windows
- Train: ~500.000 synchronized LiDAR, IMU, skeleton poses, ~37% with text descriptions

Evaluate with 3 Tasks

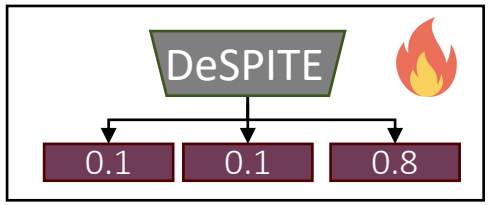
(a) Matching (RE-ID)



(b) Temporal Moment Retrieval



(c) HAR Finetuning

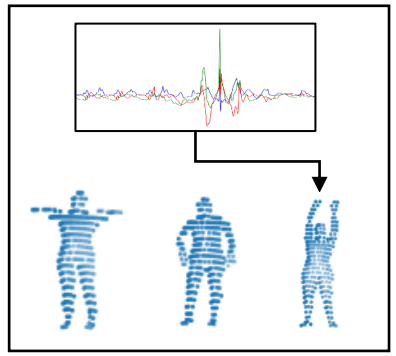


Ablations over all modality combinations in joint embeding space

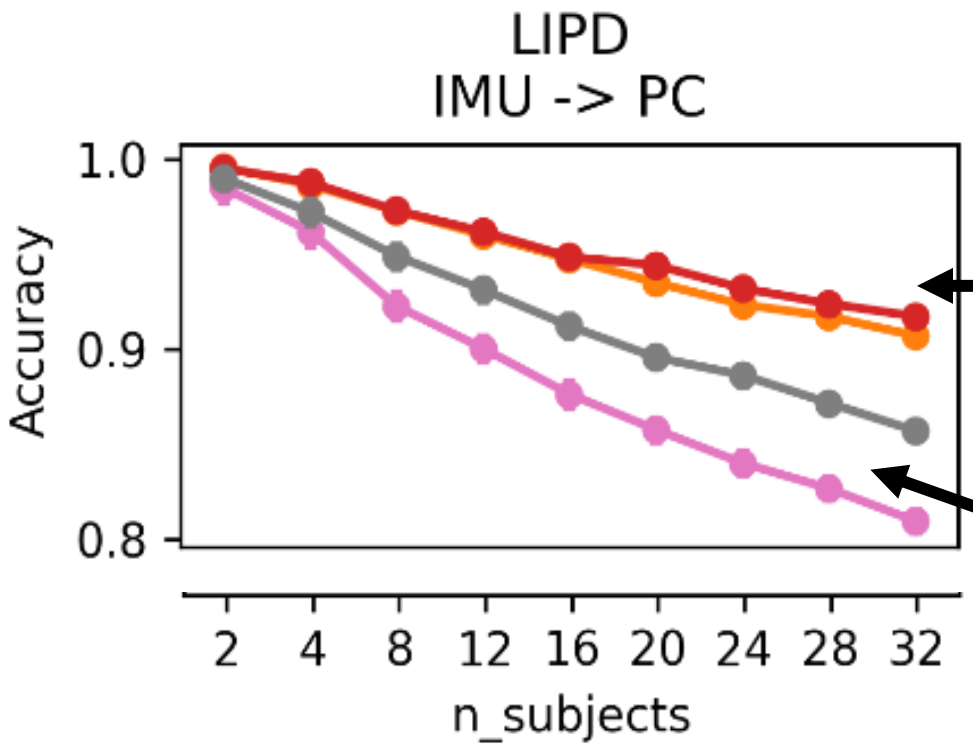
- **E.g., DeSPITE = all 4 = (Skeleton, Point Cloud, IMU, Text)**
- **E.g., DeSPIE = only 3 = (Skeleton, Point Cloud, IMU)**
- **E.g., DePIE = only 2 = (Point Cloud, IMU)**

Evaluation: Matching/RE-ID

- 3 Test Datasets (LIPD,TC,DIP)
- N=2,4, 8, ... 32 Subjects
  - 1000 random sampled scenes
- Metric: Matching Accuracy



(only showing LIPD, IMU->PC, all results in paper)



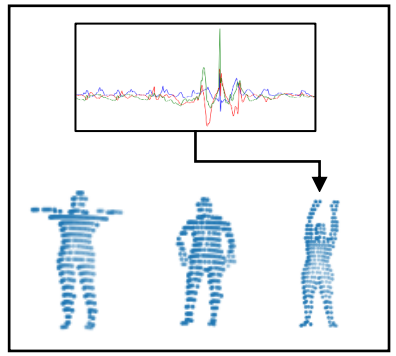
Main Findings

(1) Strong matching accuracy!
- DeSPIE
- DePIE

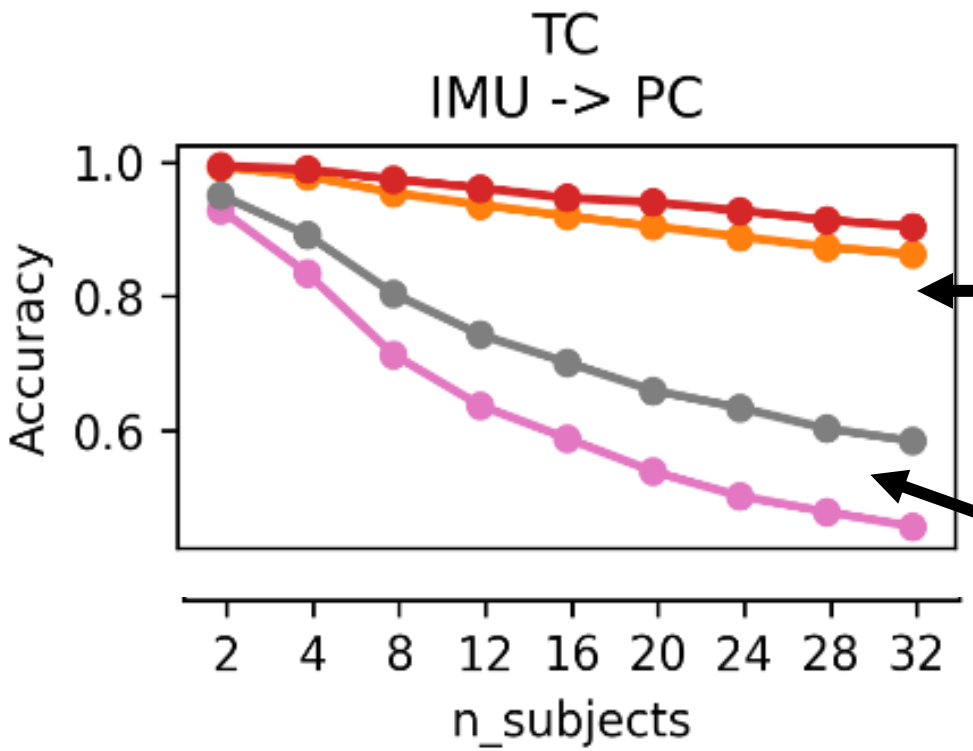
(2) Joint embeddings trained with text perform worse!
- DePITE
- DeSPITE

Evaluation: Matching/RE-ID

- 3 Test Datasets (LIPD,TC,DIP)
- N=2,4, 8, ... 32 Subjects
  - 1000 random sampled scenes
- Metric: Matching Accuracy



(only showing TC, IMU->PC, all results in paper)



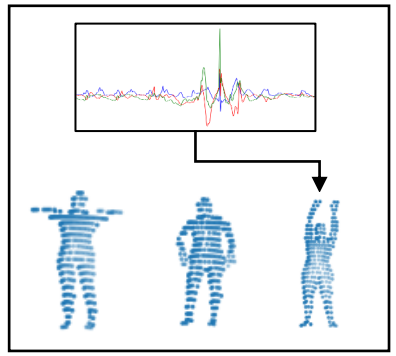
Main Findings

(1) Strong matching accuracy!
- DeSPIE
- DePIE

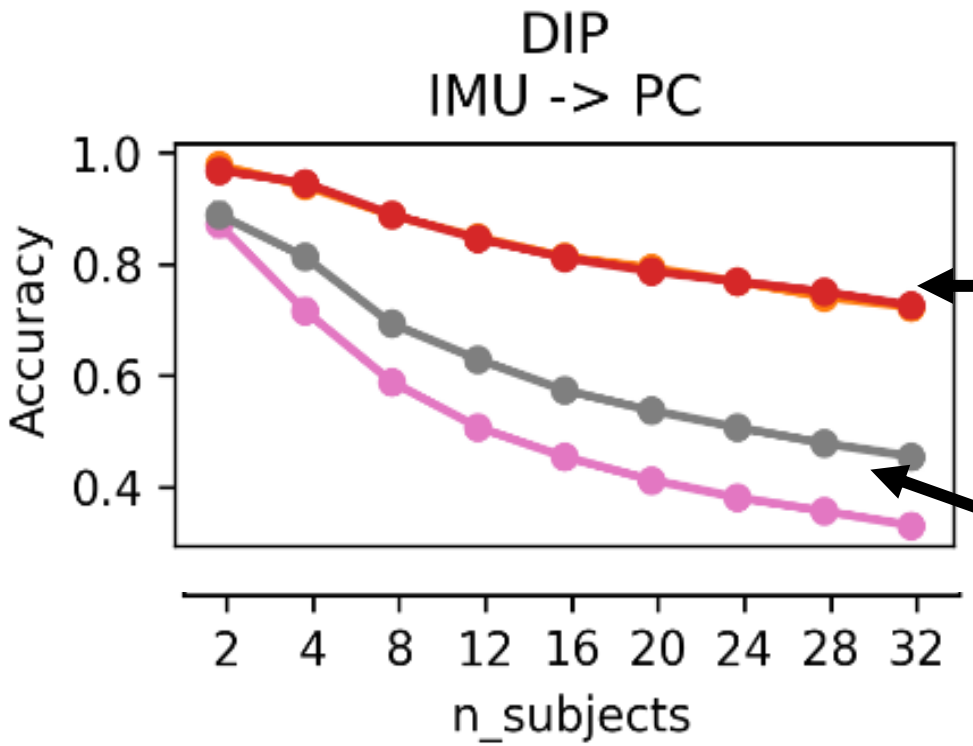
(2) Joint embeddings trained with text perform worse!
- DePITE
- DeSPITE

Evaluation: Matching/RE-ID

- 3 Test Datasets (LIPD,TC,DIP)
- N=2,4, 8, ... 32 Subjects
  - 1000 random sampled scenes
- Metric: Matching Accuracy



(only showing DIP, IMU->PC, all results in paper)

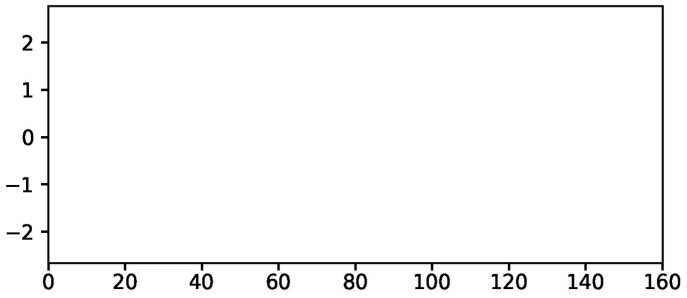


Main Findings

(1) Strong matching accuracy!
- DeSPIE
- DePIE

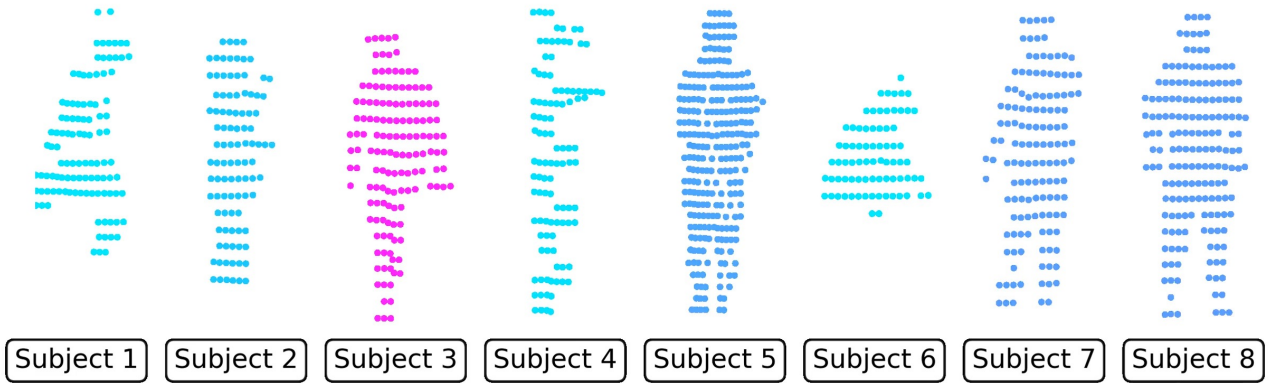
(2) Joint embeddings trained with text perform worse!
- DePITE
- DeSPITE

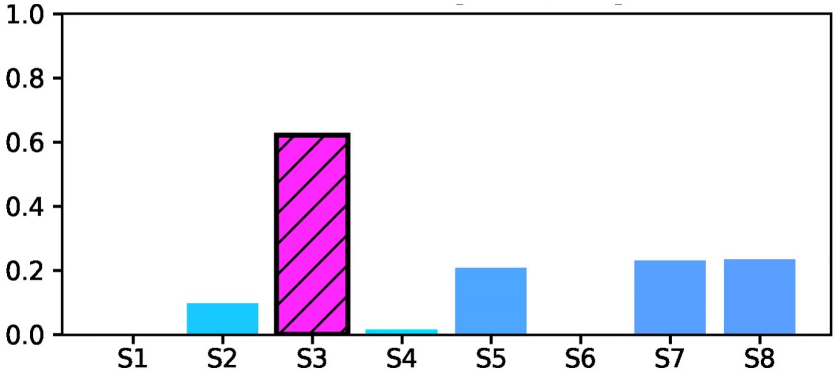
Evaluation: Matching/RE-ID



IMU Query

Candidates in Multi-Person Scene

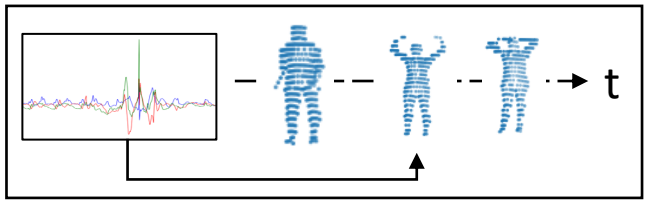




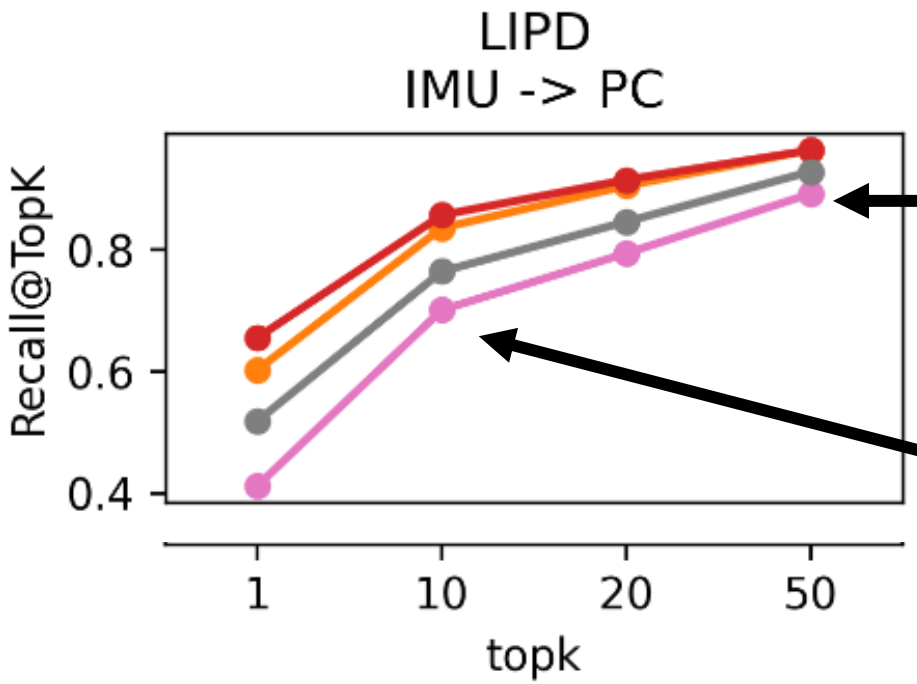
Similarity scores to each candidate

Evaluation: Temporal Moment Retrieval

- 3 Test Datasets (LIPD, TC, DIP)
- Find a subsequence in full sequence
- Metric: Recall@Top1, 10, 20, 50
  → Correct if within 1 second



(only showing IMU->PC, all combinations in paper)



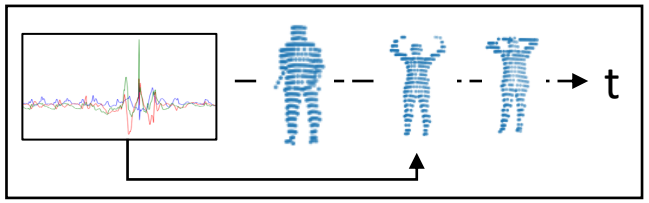
Main Findings

(1) Strong moment retrieval!
•DeSPIE •DePIE

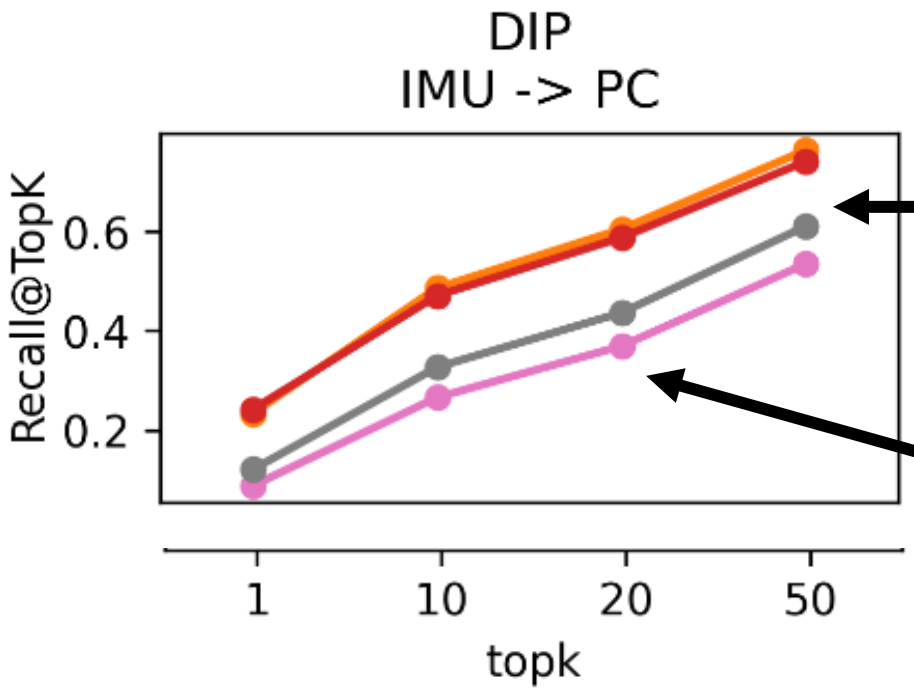
(2) Joint embeddings trained with text perform worse!
•DePITE •DeSPITE

Evaluation: Temporal Moment Retrieval

- 3 Test Datasets (LIPD, TC, DIP)
- Find a subsequence in full sequence
- Metric: Recall@Top1, 10, 20, 50
  → Correct if within 1 second



(only showing IMU->PC, all combinations in paper)



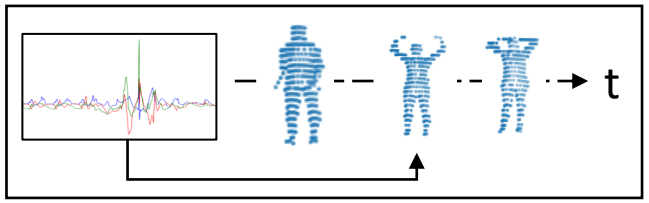
Main Findings

(1) Strong moment retrieval!
•DeSPIE •DePIE

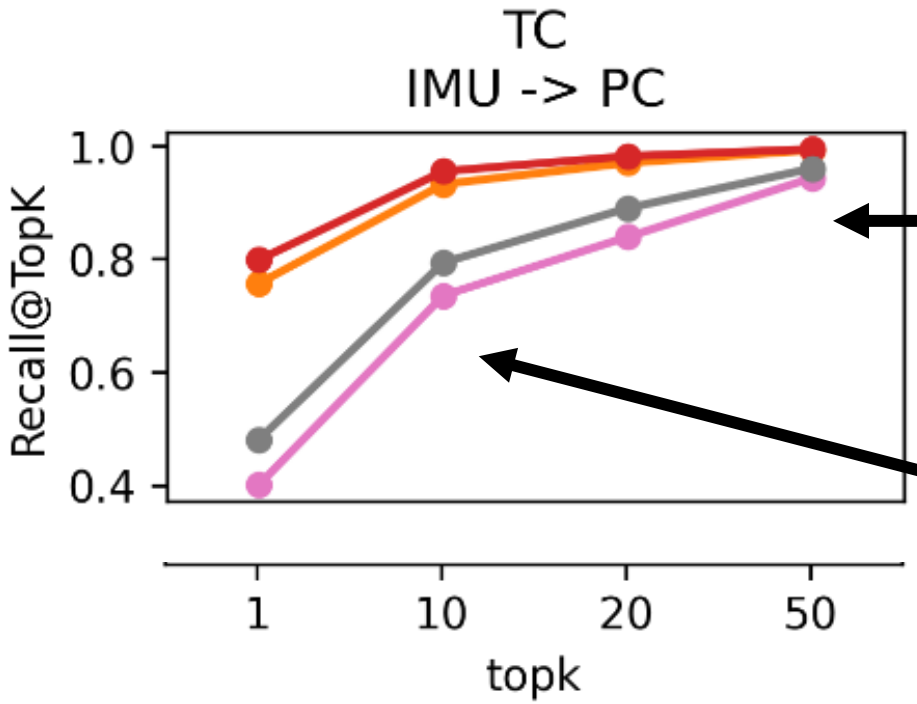
(2) Joint embeddings trained with text perform worse!
•DePITE •DeSPITE

Evaluation: Temporal Moment Retrieval

- 3 Test Datasets (LIPD, TC, DIP)
- Find a subsequence in full sequence
- Metric: Recall@Top1, 10, 20, 50
  → Correct if within 1 second



(only showing IMU->PC, all combinations in paper)



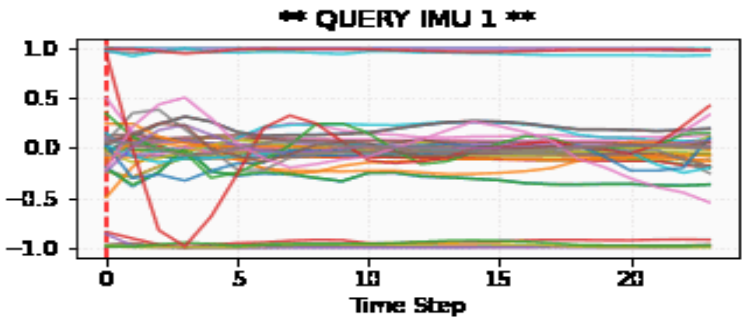
Main Findings

(1) Strong moment retrieval!
•DeSPIE •DePIE

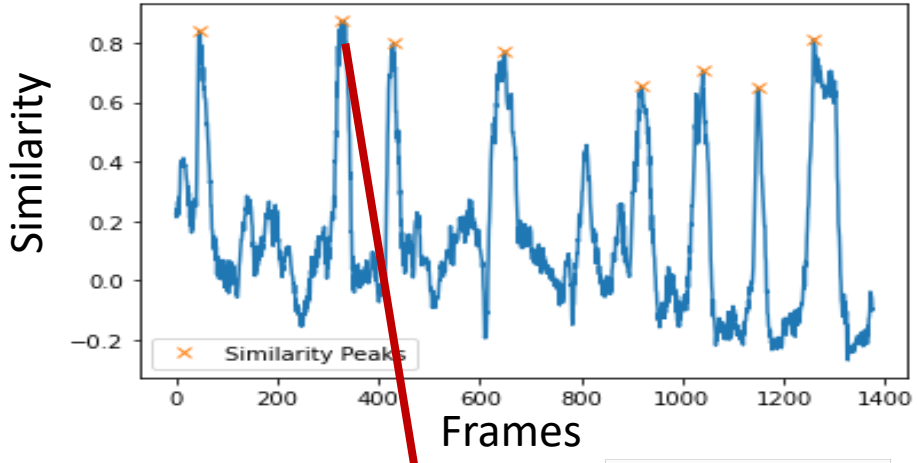
(2) Joint embeddings trained with text perform worse!
•DePITE •DeSPITE

Evaluation: Temporal Moment Retrieval

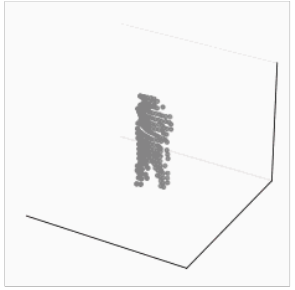
IMU Represents: ?



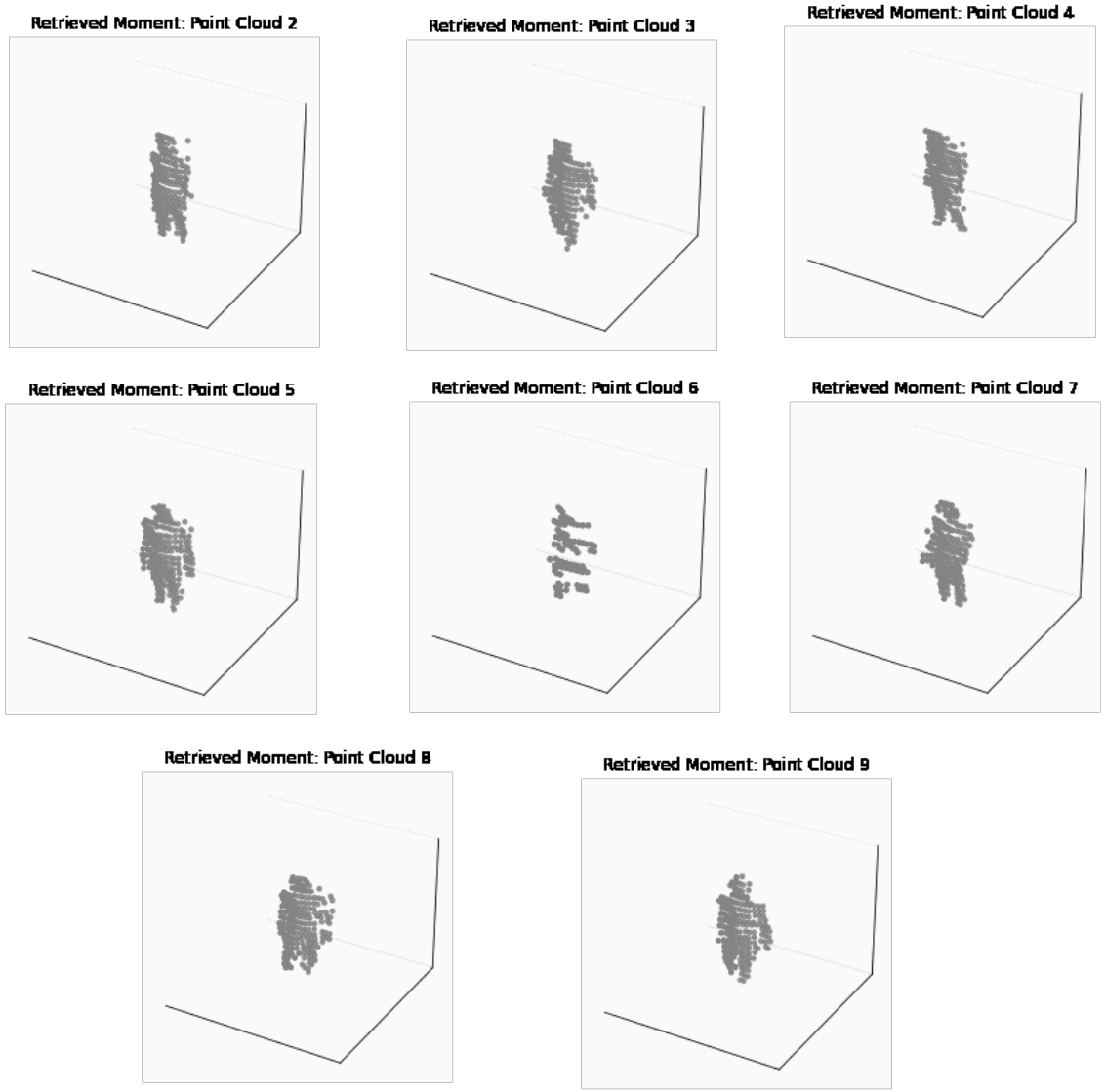
Compute similarity to each frame in point cloud video



Ground Truth

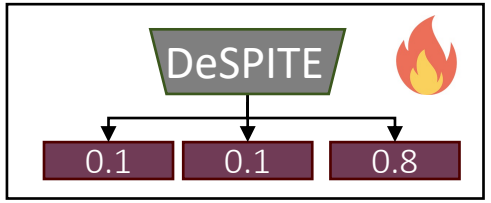


Most similar moments in point cloud video



Evaluation: Point Cloud HAR finetuning

Datasets:
(1) HMPEAR[1]
(2) MSR-Action3D[2]
Metric: Accuracy



HMPEAR:

Method	Modality	Acc(Seg)↑
Uni-Modal Supervised Learning Only		
PSTNet [11]	PC	64.3
P4-Transformer [10]	PC	63.9
PST-Transformer† [13]	PC	65.94
Multi-Modal Supervised Learning Only		
AR-Proj [29]	RGB+PC	60.6
PEAR-Proj (BestPE) [29]	RGB+PC	64.1
PEAR-Proj (BestAR) [29]	RGB+PC	66.0
Multi-Modal Pre-Training + Transfer Learning		
PST-Transformer + **DeSPITE (ours)**	PC	**69.18** *(+3.24)*
PST-Transformer + **DeSPIE (ours)**	PC	**70.26** *(+4.32)*
PST-Transformer + **DePITE (ours)**	PC	**70.65** *(+4.71)*

MSR-Action3D:

Methods	Video Acc@1 (↑)
Supervised Learning Only	
PST-Transformer† [13]	92.33
MAMBA4D [31]	93.38
PvNext [58]	94.77
KAN-HyperpointNet [7]	95.59
Uni-Modal Pre-Training + Transfer Learning	
PSTNet + PointCPSC [52]	92.68 (+1.48)
PSTNet + PointCMP [51]	93.27 (+2.07)
PST-Transformer + MaST-Pre [50]	94.08 (+0.35)
PPTr + C2P [63]	94.76 (+2.43)
P4Transformer + M2PSC [18]	93.03 (+2.09)
PST-Transformer + M2PSC [18]	94.84 (+1.11)
Multi-Modal Pre-Training + Transfer Learning	
PST-Transformer + **DePITE (Ours)**	**95.12** (+1.39 | + 2.79†)
PST-Transformer + **DeSPIE (Ours)**	**95.47** (+1.74 | + 3.14†)
PST-Transformer + **DeSPITE (Ours)**	**95.47** (+1.74 | + 3.14†)

Main Findings:

- **Multi-Modal Pre-Training better than Uni-Modal Pre-Training**
- **Embeddings trained with text or all modalities benefits HAR (DeSPIE, DePITE, DeSPITE always best)**

1 Lin et al., 2024, Hmpear: A dataset for human pose estimation and action recognition
2 Li et al., 2010, Action recognition based on a bag of 3d points

ICCV HONOLULU HAWAII
OCT 19-23, 2025

TECHNISCHE UNIVERSITÄT DARMSTADT

DeSPITE: Exploring Contrastive Deep Skeleton-Pointcloud-IMU-Text Embeddings for Advanced Point Cloud Human Activity Understanding

Thank you for listening.

Telecooperation Lab