ISMAR 2025

HaHeAE: Learning Generalisable Joint Representations of Human Hand and Head Movements in Extended Reality

Zhiming Hu[1,2], Guanhua Zhang[1], Zheming Yin[1], Daniel Häufle[3,4], Syn Schmitt[1,4], Andreas Bulling[1]

[1]University of Stuttgart

[2]The Hong Kong University of Science and Technology (Guangzhou)

[3]University of Tuebingen

[4]The Center for Bionic Intelligence Tuebingen Stuttgart

zhiminghu.net/hu25_haheae

Table of Contents

Applications of human hand and head movements in XR

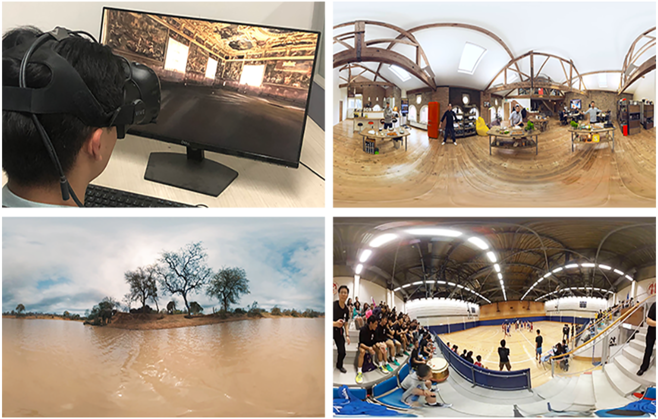
Interaction target prediction [Belardinelli IROS'22]

Research Background

Applications of human hand and head movements in XR

Redirected walking [Gandrud SAP'16]

Applications of human hand and head movements in XR



Activity recognition [Hu TVCG'22]

Applications of human hand and head movements in XR

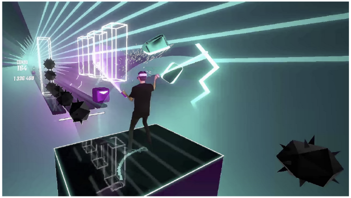


Figure 1: “Beat Saber” – VR rhythm game.

Figure 2: “Tilt Brush” – VR painting app.

User identification [Nair TVCG’24]

Learning **generalisable joint representations** of human hand and head movements in XR

- **Jointly** modelling hand and head movements in XR has significant potential for understanding human behaviours
- **Generalisable** hand-head representations can be reused for various XR applications

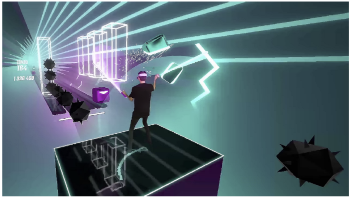
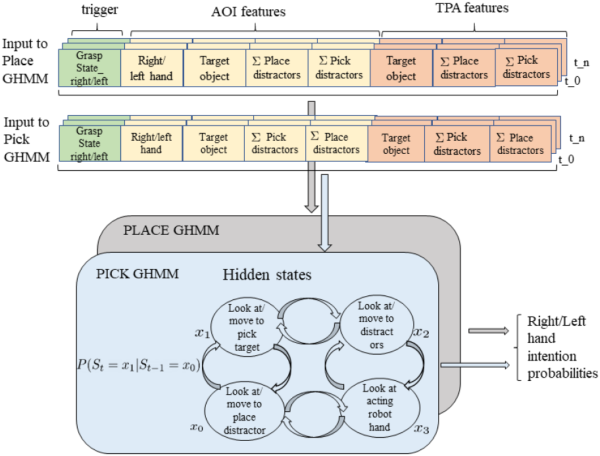


Figure 1: "Beat Saber" – VR rhythm game.

Figure 2: "Tilt Brush" – VR painting app.

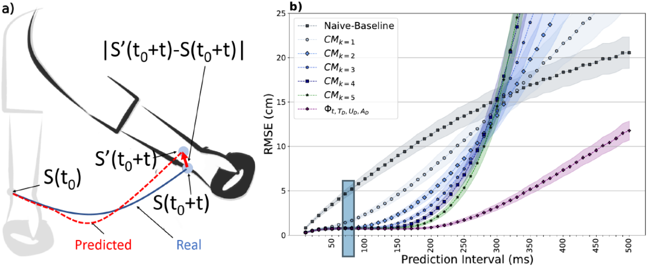
Table of Contents

Hand behaviour modelling



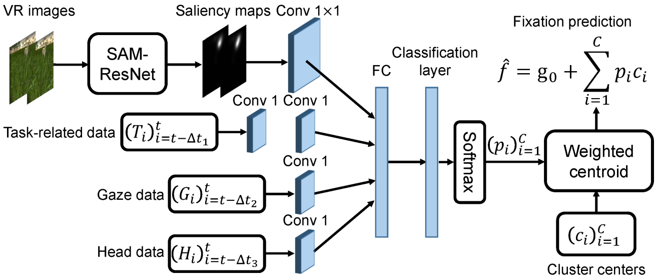
Gaussian hidden Markov models for intention estimation [Belardinelli IROS'22]

Hand behaviour modelling



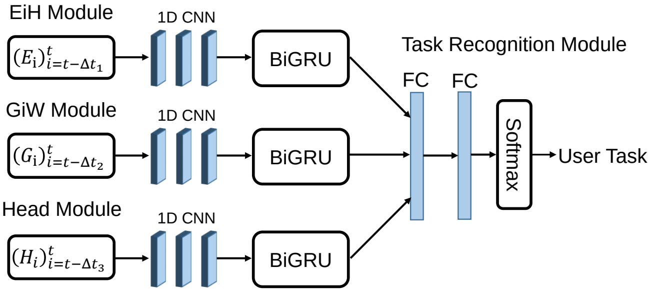
Kinematic regressive model for hand trajectory prediction [Gamage UIST'21]

Head behaviour modelling



1D CNN for fixation prediction [Hu TVCG'21]

Head behaviour modelling



1D CNN and BiGRU for activity recognition [Hu TVCG'22]

Related Work: Hand and Head Behaviour Modelling

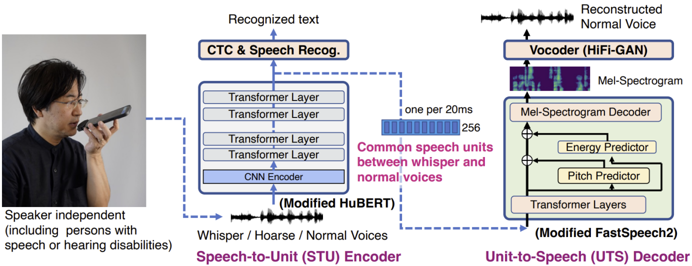
Previous works

- Only focus on a **single** modality (hand or head)
- Limited to a **specific** XR application

Our work

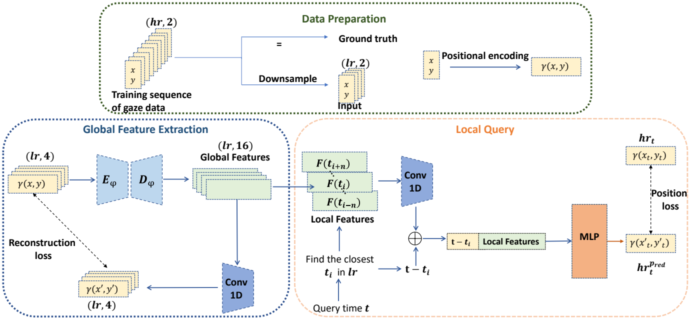
- **Jointly** modelling hand and head behaviours
- **Generalisable** representations for various XR applications

Learning generalisable representations of speech signals



Transformer-based autoencoder for speech signals [Rekimoto CHI'23]

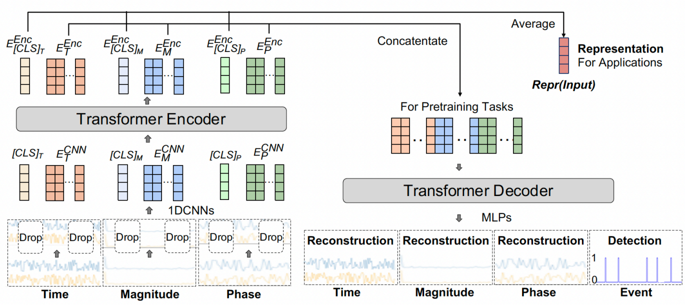
Learning generalisable representations of gaze behaviour



Implicit neural representation learning for gaze data [Jiao UIST’23]

Related Work: Generalisable Representation Learning

Learning generalisable representations of mouse behaviour



Transformer-based autoencoder for mouse behaviour [Zhang CHI'24]

Previous works

- Learning generalisable representations of **speech, gaze, or mouse** behaviours

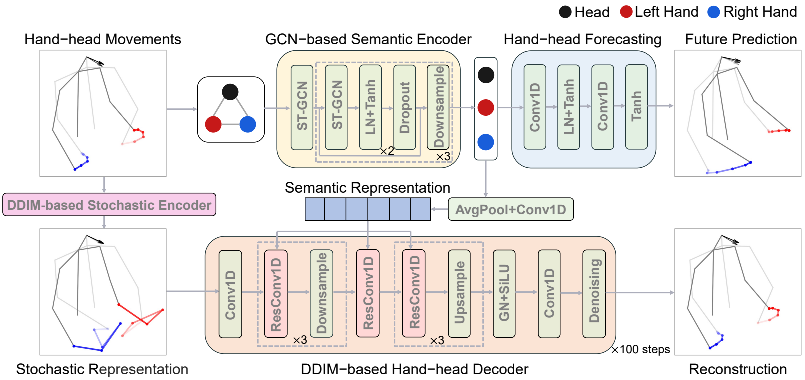
Our work

- Learning generalisable representations of **hand and head** behaviours

Table of Contents

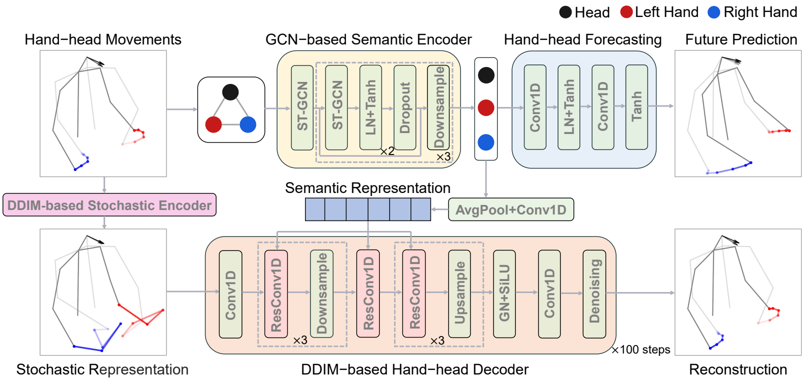
Problem formulation

- Given a sequence of hand trajectories and head orientations
- Generate a joint semantic representation of the signals



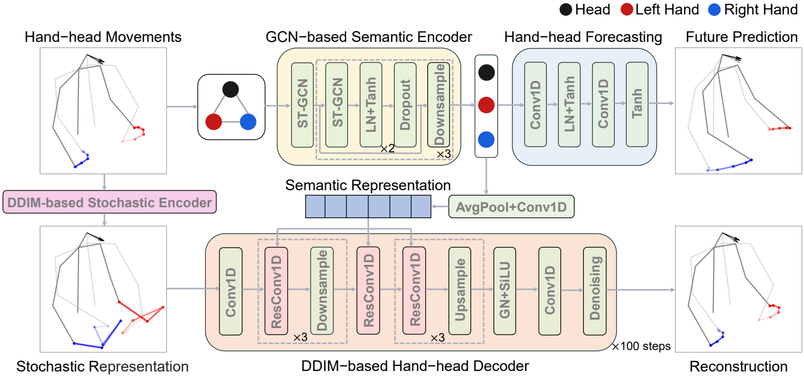
GCN-based semantic encoder

- Treat hand and head as **nodes** in a graph
- Spatio-temporal **GCN** for learning **semantic** representation



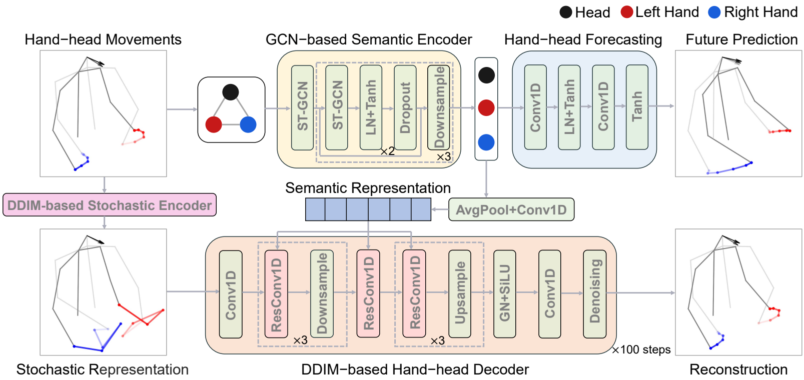
DDIM-based stochastic encoder

- **DDIM**-based encoder for learning **stochastic** representation



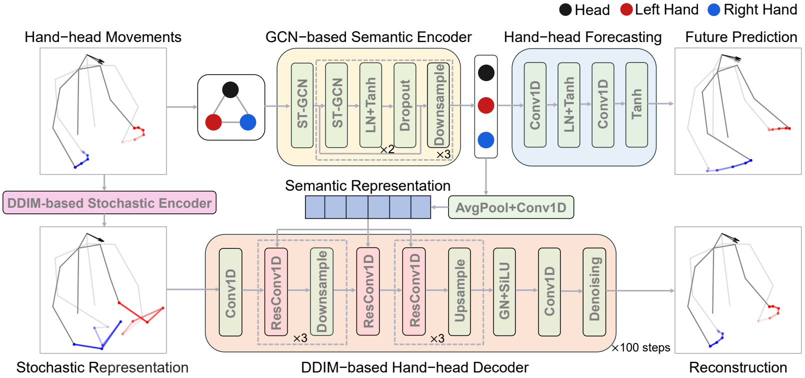
DDIM-based hand-head decoder

- Use semantic representation as a **condition** to **DDIM**
- Use DDIM to **reconstruct** the original hand-head movements



Hand-head forecasting

- **Auxiliary training task** to refine the semantic representation



Reconstruction evaluation settings

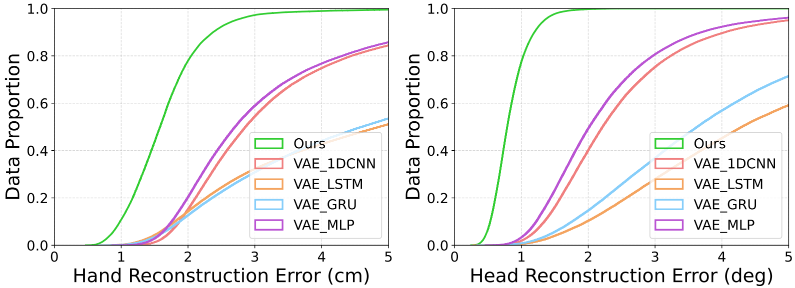
- Training: **EgoBody** [Zhang ECCV'22] dataset
- Test: **ADT** [Pan ICCV'23] and **GIMO** [Zheng ECCV'22] datasets
- Metric for hand reconstruction: mean position error (cm)
- Metric for head reconstruction: mean angular error (deg)
- Sequence length: 40 frames
- Baselines: VAE_1DCNN, VAE_LSTM, VAE_GRU, VAE_MLP

Reconstruction performance

	EgoBody		ADT		GIMO	
	hand	head	hand	head	hand	head
VAE_1DCNN	3.575	2.549	3.876	2.776	4.422	3.100
VAE_LSTM	7.254	5.421	7.933	9.928	8.908	8.060
VAE_GRU	6.776	4.390	7.351	6.161	8.369	6.371
VAE_MLP	3.455	2.338	3.932	2.733	4.310	2.927
Ours	**1.664**	**0.834**	**1.966**	**0.707**	**2.397**	**1.247**

Our method **significantly outperforms** other methods for reconstructing both hand and head signals

Reconstruction performance



(a) EgoBody (Hand) (b) EgoBody (Head)

Our method achieves **significantly better** performance than other methods in terms of reconstruction error distributions

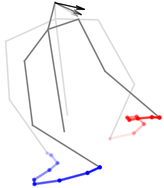
Ablation study

	EgoBody		ADT		GIMO	
	hand	head	hand	head	hand	head
Ours_1DCNN	2.010	1.070	2.370	1.095	2.883	1.500
Ours_LSTM	1.706	0.842	**1.937**	0.713	2.587	1.341
Ours_GRU	1.715	0.861	1.964	0.718	2.658	1.377
Ours_MLP	1.840	0.851	2.213	0.822	2.660	1.279
Ours w/o E_{sem}	56.803	93.361	55.952	92.991	57.005	91.864
Ours w/o E_{sto}	10.604	11.278	11.889	12.878	11.158	12.042
Ours	**1.664**	**0.834**	1.966	**0.707**	**2.397**	**1.247**

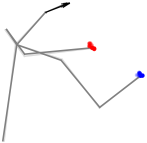
Each component contributes to our method's performance

Table of Contents

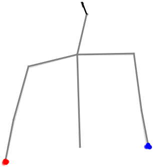
Representative hand-head clusters and their semantics



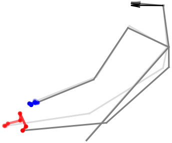
(a) Activity: **Instruct to act**. The head is facing slightly downward; both hands move noticeably below the head.



(b) Activity: **Learn course while sitting**. The head is facing slightly upward; both arms are bent and have no large movements.



(c) Activity: **Casually chat while standing**. The head is facing upward; both arms are laid down and remain almost still.



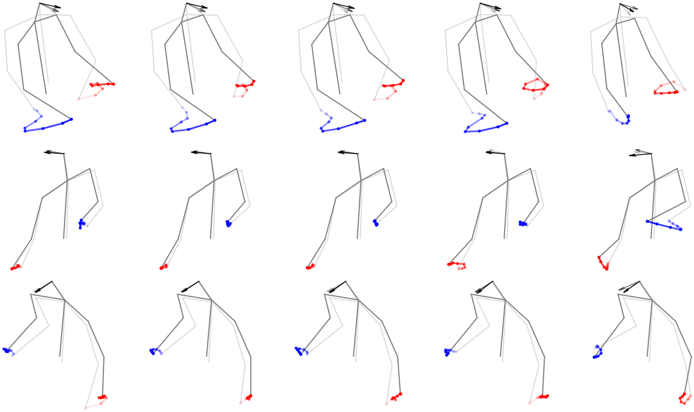
(d) Activity: **Take a tape**. The head is facing forward; the left hand has a greater range of motion than the right hand.

Clustering performance of different methods

	DBI ↓	CHI ↑
VAE_1DCNN	2.182	19.582
VAE_LSTM	1.343	26.626
VAE_GRU	1.367	28.892
VAE_MLP	1.837	29.485
Ours	**1.141**	**37.970**

Our method outperforms other methods in clustering performance

Generation results



Original | E_{sto} | $E_{sto} + 0.1 \cdot z$ | $E_{sto} + 0.3 \cdot z$ | z

Our method can be used to generate variable hand-head data

Performance on downstream tasks

	User Identification	Activity Recognition	
	EgoBody	EgoBody	ADT
Chance	8.3%	33.3%	33.3%
VAE_1DCNN	26.3%	48.8%	62.7%
VAE_LSTM	24.9%	47.1%	61.3%
VAE_GRU	28.0%	41.3%	61.0%
VAE_MLP	25.8%	50.5%	60.7%
Ours *hand only*	18.0%	55.5%	53.6%
Ours *head only*	25.7%	46.1%	62.5%
Ours w/o *forecasting*	29.4%	54.7%	63.1%
Ours	**29.8%**	**55.7%**	**63.9%**

Our method consistently outperforms other methods on downstream tasks

Table of Contents

Limitations

- Evaluations are limited to existing XR datasets
- Ignore the scene context information

Future work

- Evaluate for a **broader** range of activities and environments
- Explore more **applications** of hand-head joint representations
- Learn joint representations of **hand**, **head**, and **scene context**

Main contributions

- A novel **representation learning** method that contains a **GCN-based** semantic encoder, a **diffusion-based** stochastic encoder, and a **diffusion-based** hand-head decoder
- Extensive experiments on **three public XR datasets** that demonstrate the effectiveness of our method
- Experiments on three use cases (**clustering**, **generation**, and **downstream tasks**) that validate our method's usefulness

Belardinelli IROS'22. Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In *Proceedings of the 2022 IEEE International Conference on Intelligent Robots and Systems*, pages 9806–9813. IEEE, 2022.

Gamage UIST'21. So predictable! continuous 3d hand trajectory prediction in virtual reality. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 332–343, 2021.

Gandrud SAP'16. Predicting destination using head orientation and gaze direction during locomotion in vr. In *Proceedings of the 2016 ACM Symposium on Applied Perception*, pages 31–38, 2016.

Hu TVCG'21. Fixationnet: forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2681–2690, 2021.

Hu TVCG'22. Ehtask: recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

Jiao UIST'23. Supreyes: Super resolutin for eyes using implicit neural representation learning. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–13, 2023.

Nair TVCG'24. Berkeley open extended reality recordings 2023 (boxrr-23): 4.7 million motion capture recordings from 105,000 xr users. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

Pan ICCV'23. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023.

Rekimoto CHI'23. Wesper: Zero-shot and realtime whisper to normal voice conversion for whisper-based speech interactions. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–12, 2023.

Zhang CHI'24. Mouse2vec: Learning reusable semantic representations of mouse behaviour. In *Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI)*, pages 1–17, 2024. doi: 10.1145/3613904.3642141.

Zhang ECCV'22. Egobody: Human body shape, motion and social interactions from head-mounted devices. In *Proceedings of the 2022 European Conference on Computer Vision*, 2022.

Zheng ECCV'22. Gimo: Gaze-informed human motion prediction in context. In *Proceedings of the 2022 European Conference on Computer Vision*, 2022.

Thank you!

Any questions?