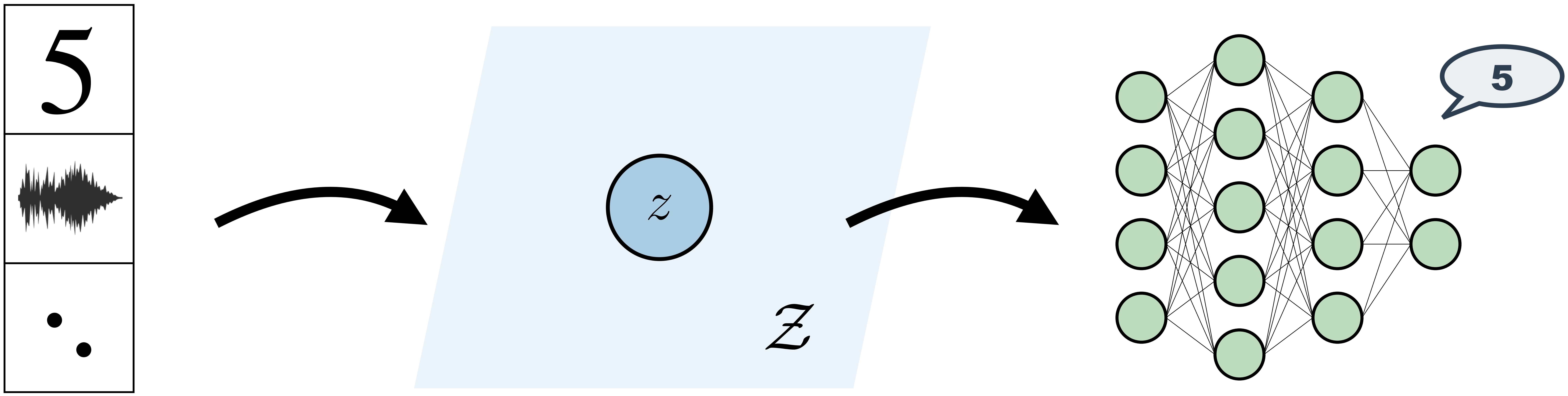
Geometric Multimodal Contrastive Representation Learning

Petra Poklukar[*,1], *Miguel Vasco*[*,2], *Hang Yin*[1], *Francisco S. Melo*[2], *Ana Paiva*[2], *Danica Kragic*[1]

[1] *KTH Royal Institute of Technology, Stockholm, Sweden*
[2] *INESC-ID & Instituto Superior Técnico, University of Lisbon, Portugal*
* *Equal contribution*

Motivation

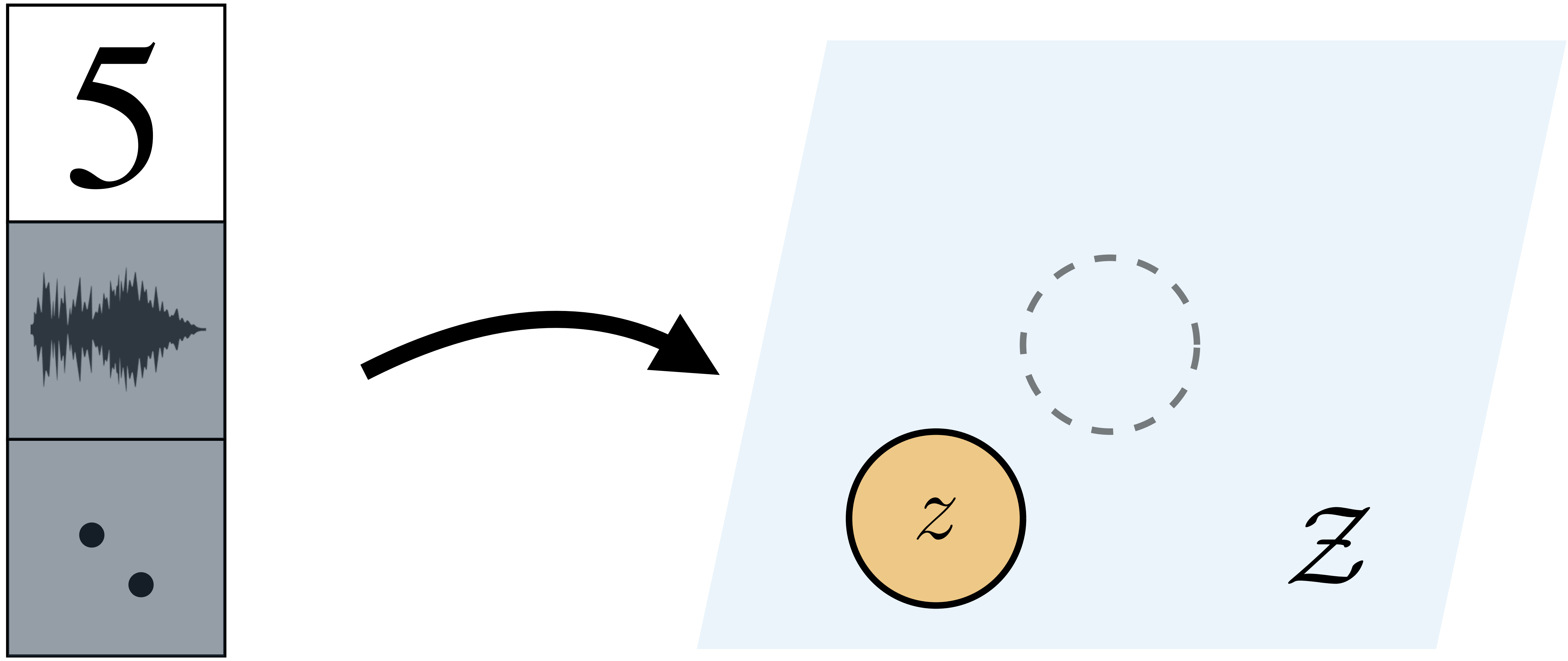


Multimodal
Observation

Latent
Representation

Downstream
Task

Motivation

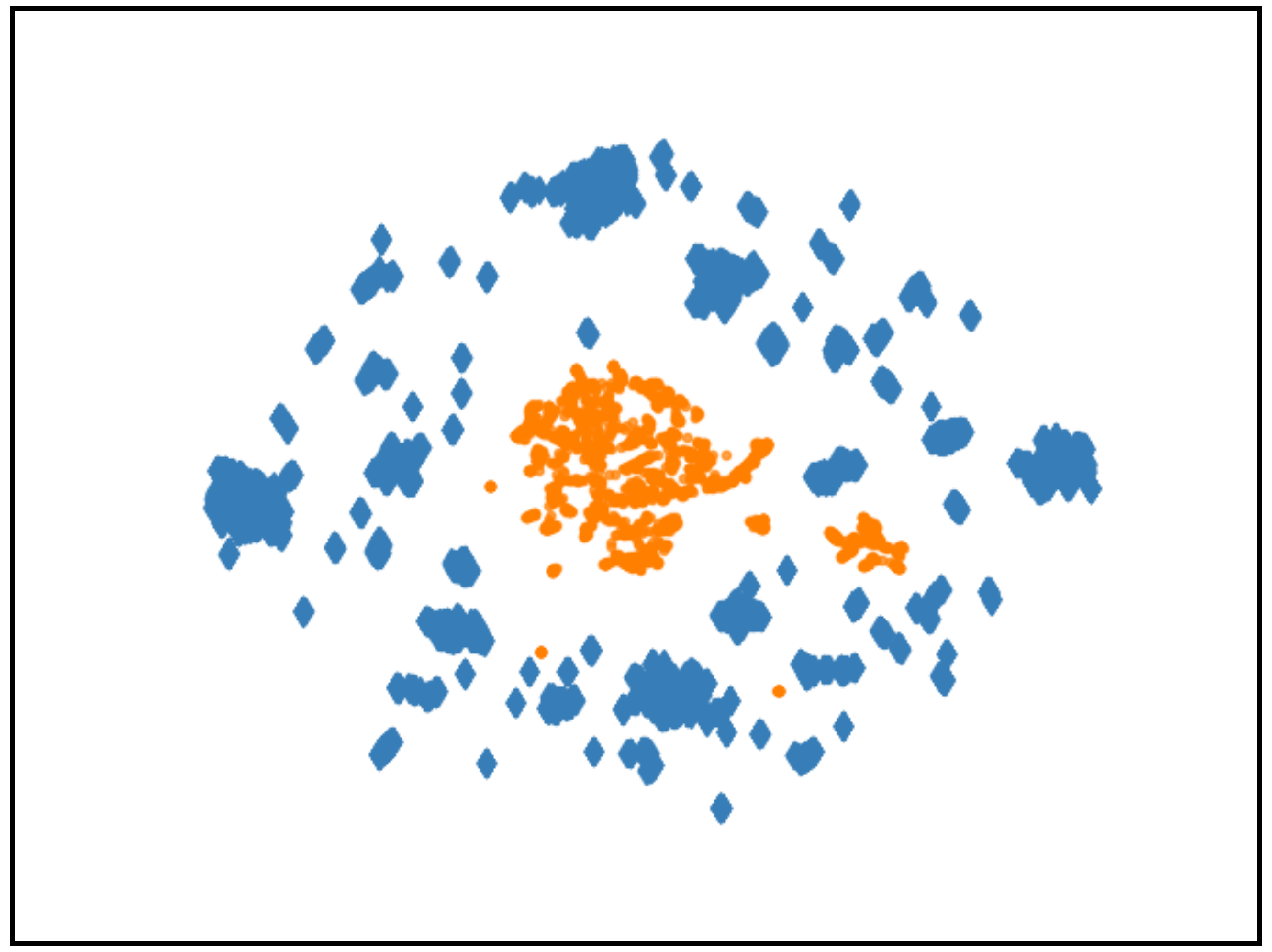


(Incomplete)
Observation

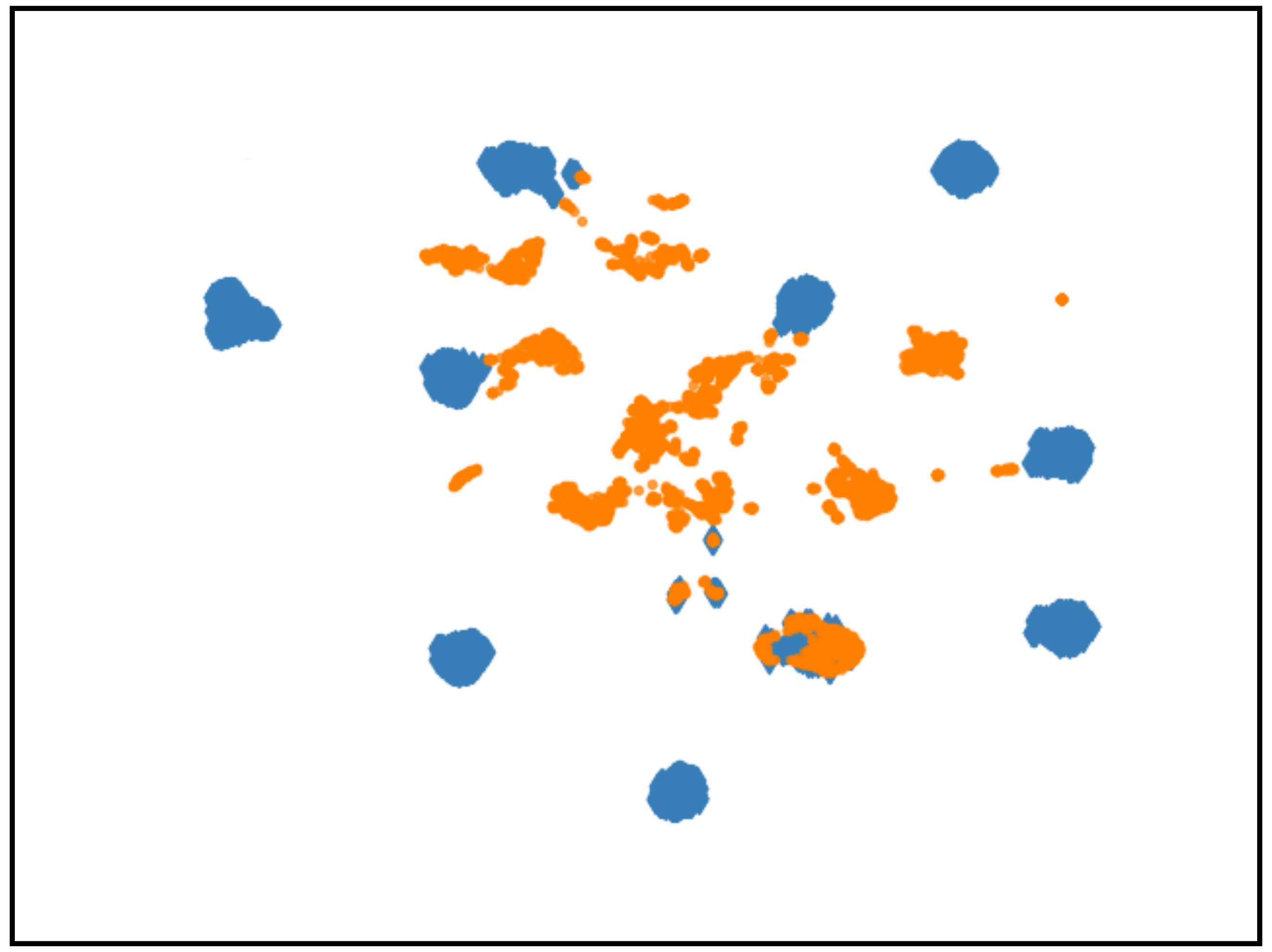
Latent
Representation

Motivation

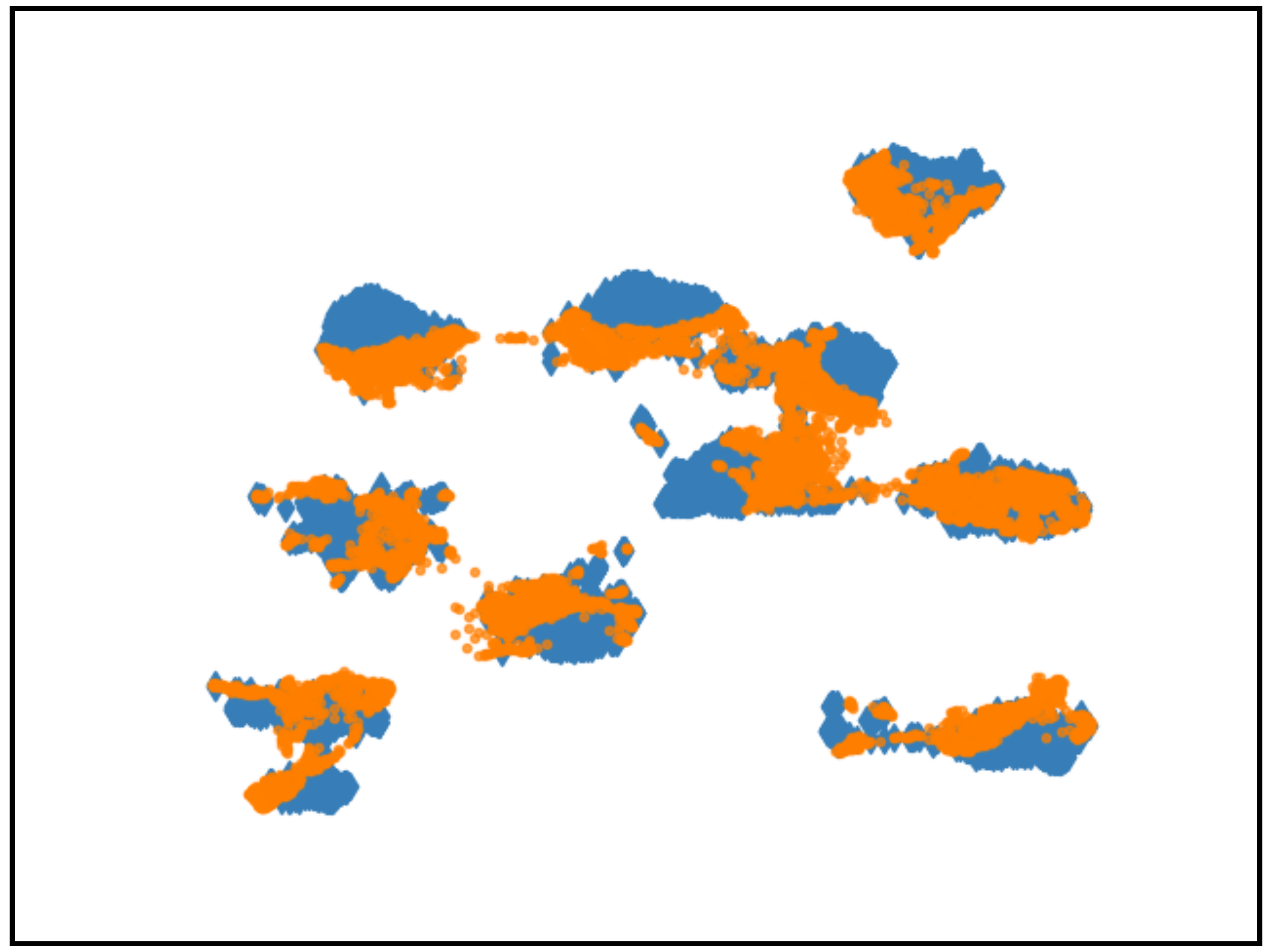
● Multimodal Observations ● Image Observations



MFM [1]



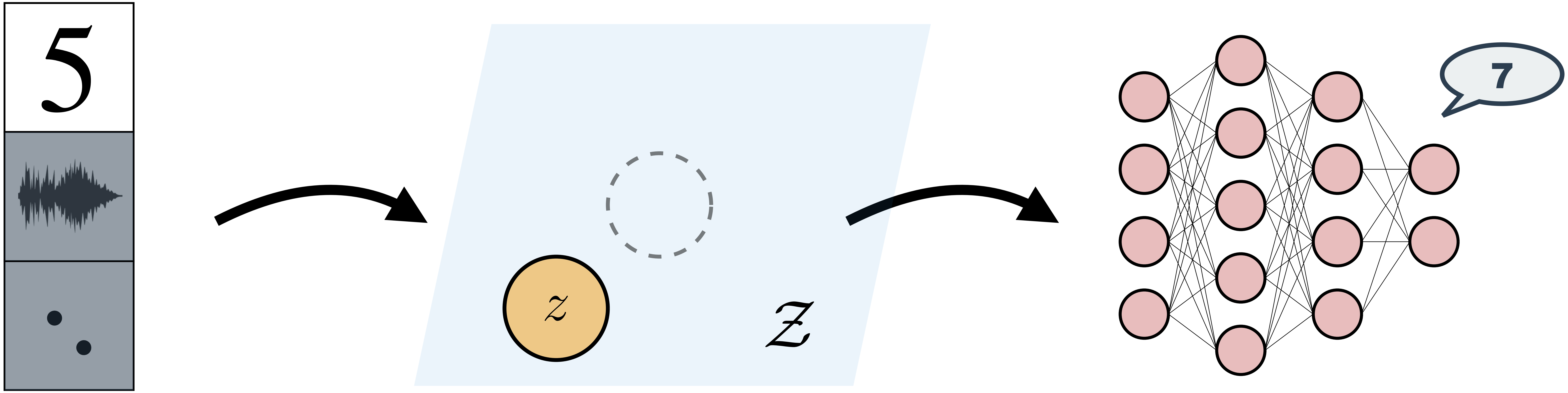
MVAE [2]



MUSE [3]

[1] Tsai, Yao-Hung Hubert, et al."Learning Factorized Multimodal Representations." *ICLR (2019)*
[2] Wu, Mike, and Noah Goodman. "Multimodal generative models for scalable weakly-supervised learning." *NeurIPS (2018)*
[3] Vasco, Miguel, et al. "How to Sense the World: Leveraging Hierarchy in Multimodal Perception for Robust Reinforcement Learning Agents." *AAMAS* (2022)

Motivation



(Incomplete)
Multimodal
Observation

Latent
Representation

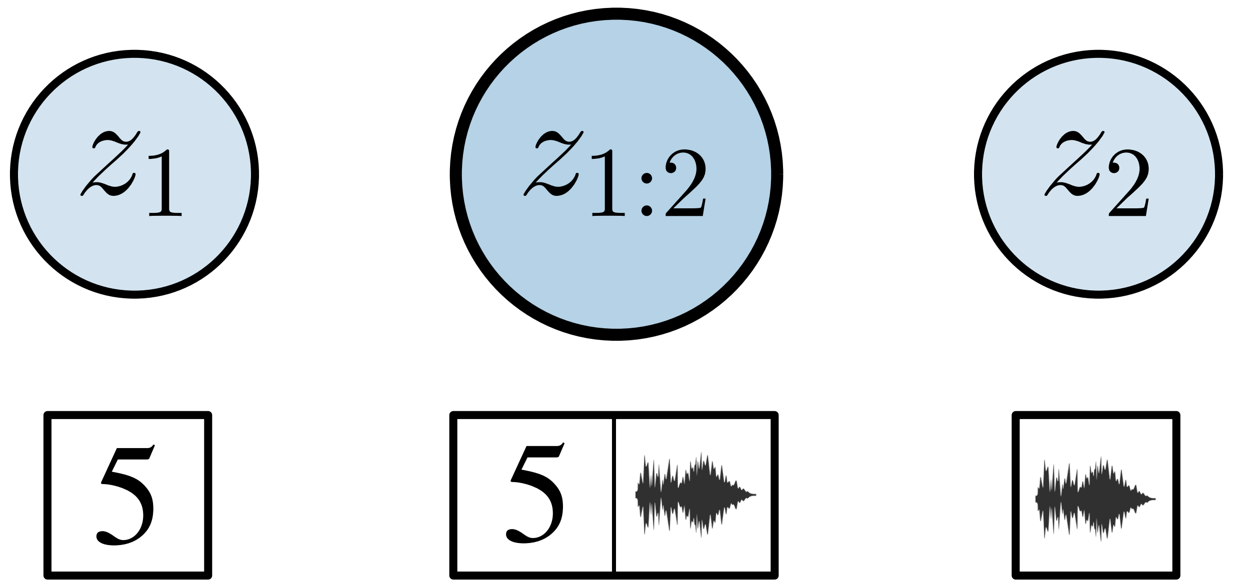
Downstream
Task

Contribution

How to learn multimodal representations for *robust* downstream performance with missing modality information?

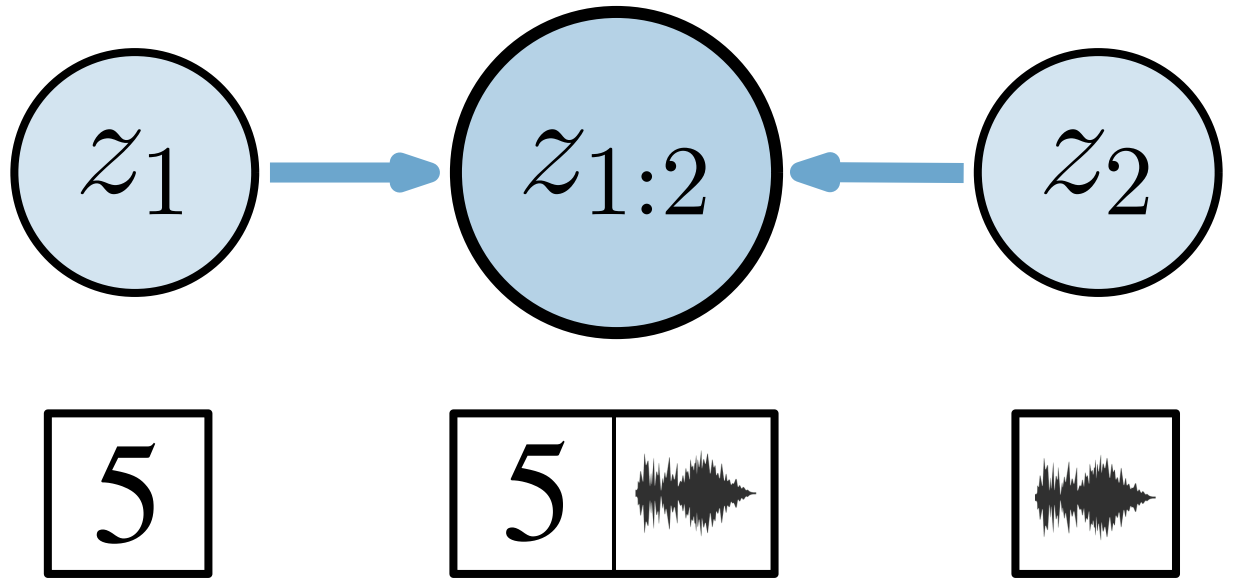
- Geometric Multimodal Contrastive (**GMC**) representation learning framework;
- **Scalable** to large number of modalities;
- **Easy to integrate** into existing architectures;
- **State-of-the-art** performance with missing modalities.

GMC: Intuition

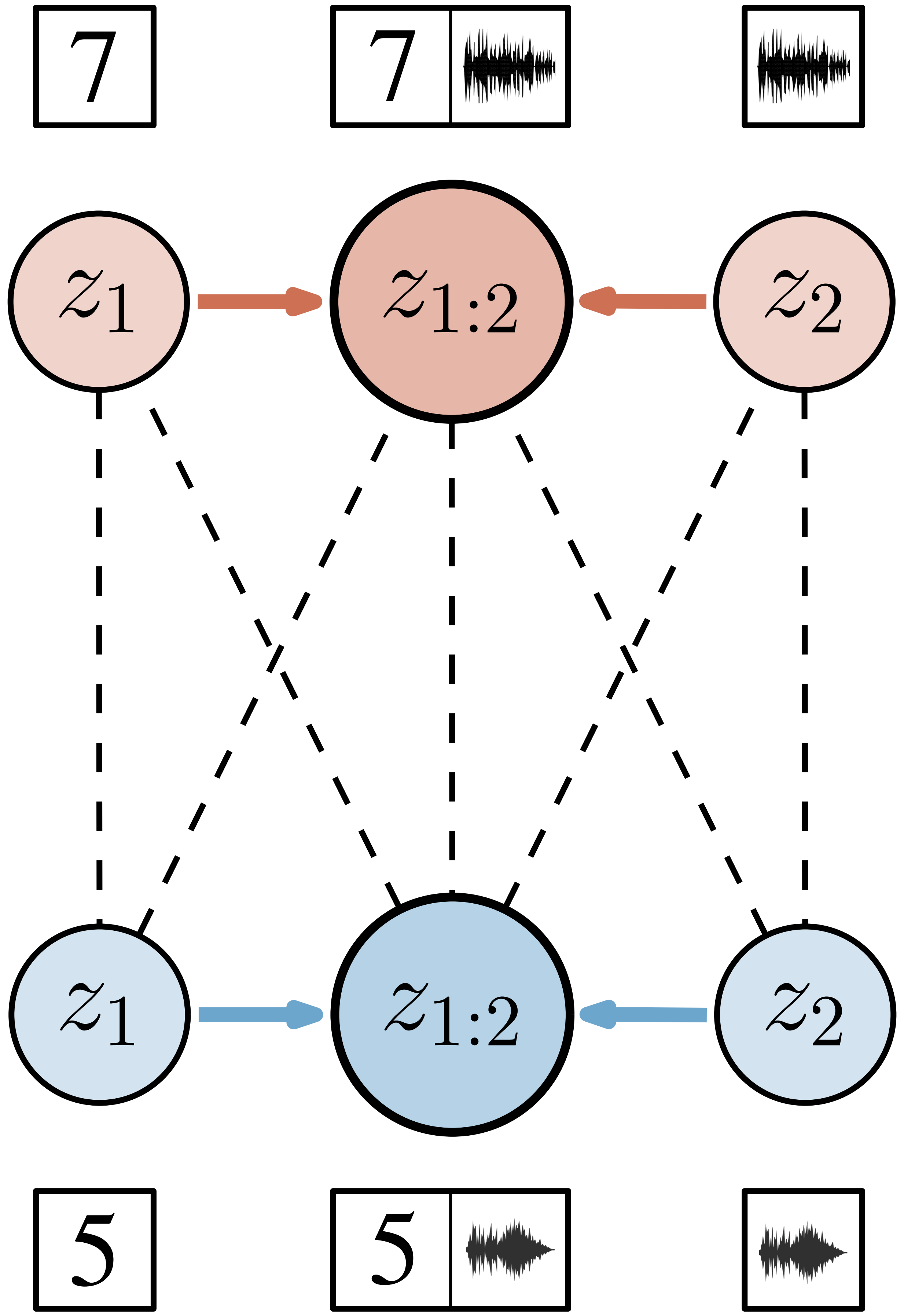


GMC: Intuition

Align complete and modality-specific representation



GMC: Intuition

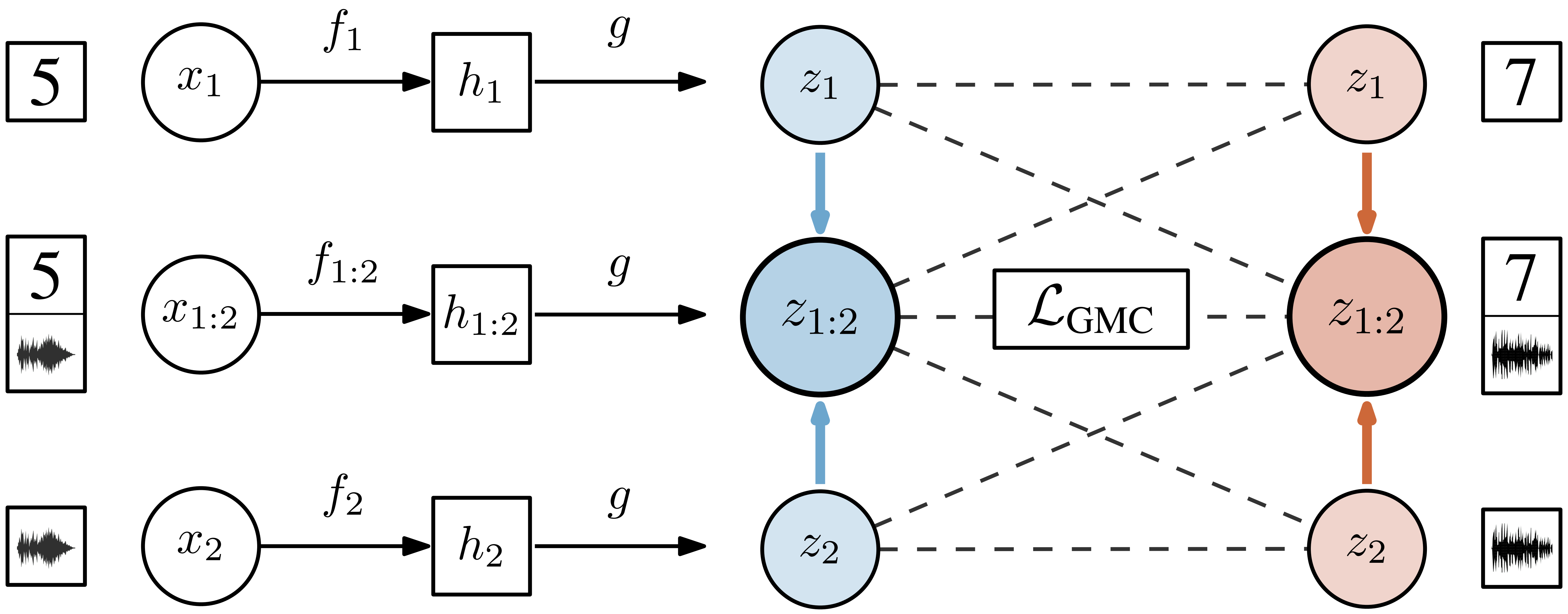


Contrast with different representations

Align complete and modality-specific representation

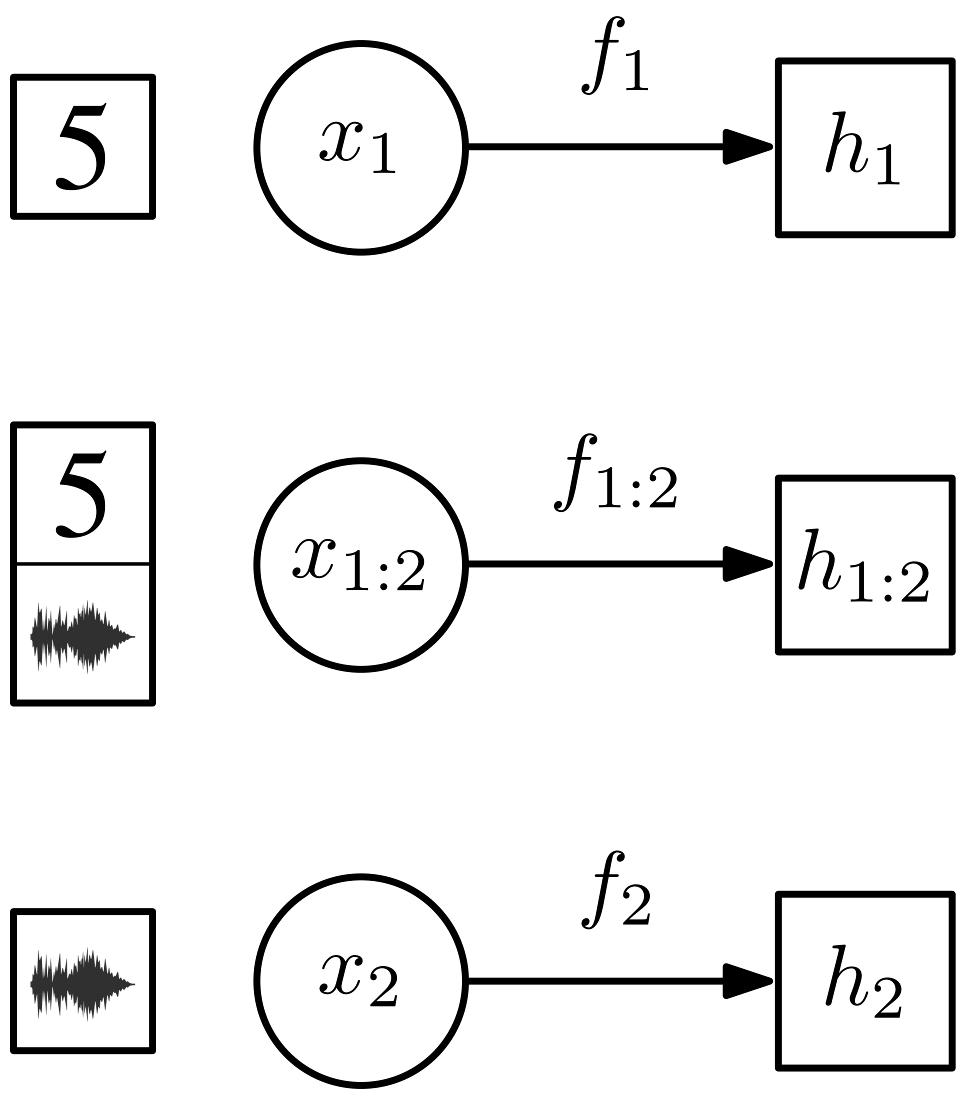
GMC: Method

Geometrical Multimodal Contrastive (GMC)



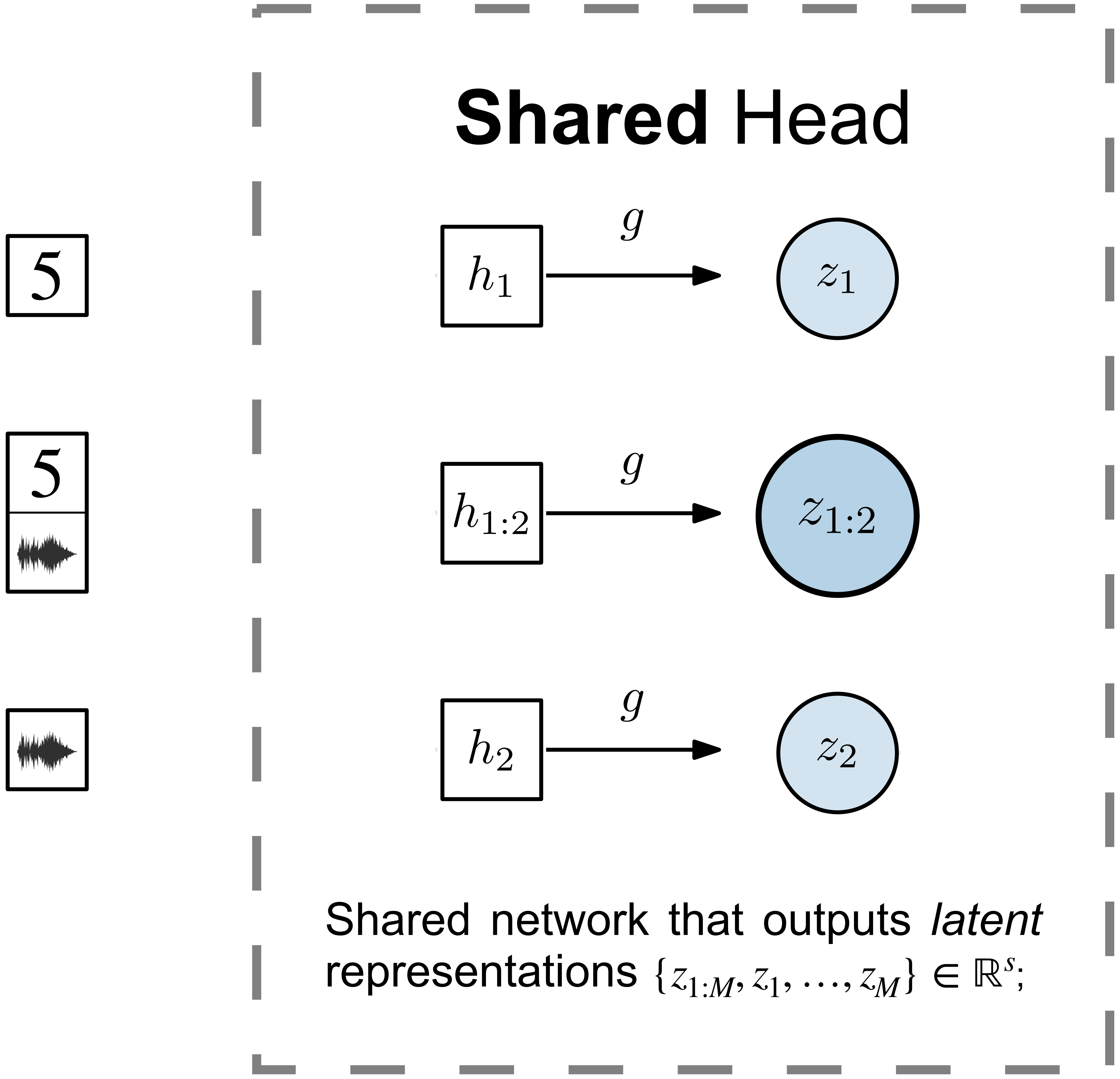
GMC: Method

Base Encoders



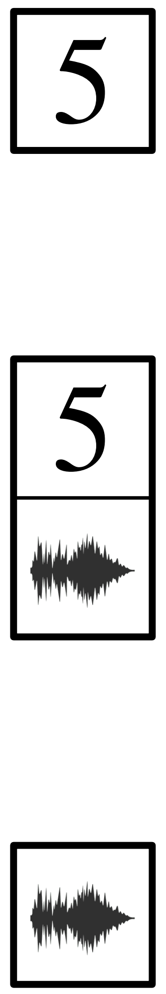
$M+1$ networks that output *intermediate* representations $\{h_{1:M}, h_1, \ldots, h_M\} \in \mathbb{R}^d$;

GMC: Method

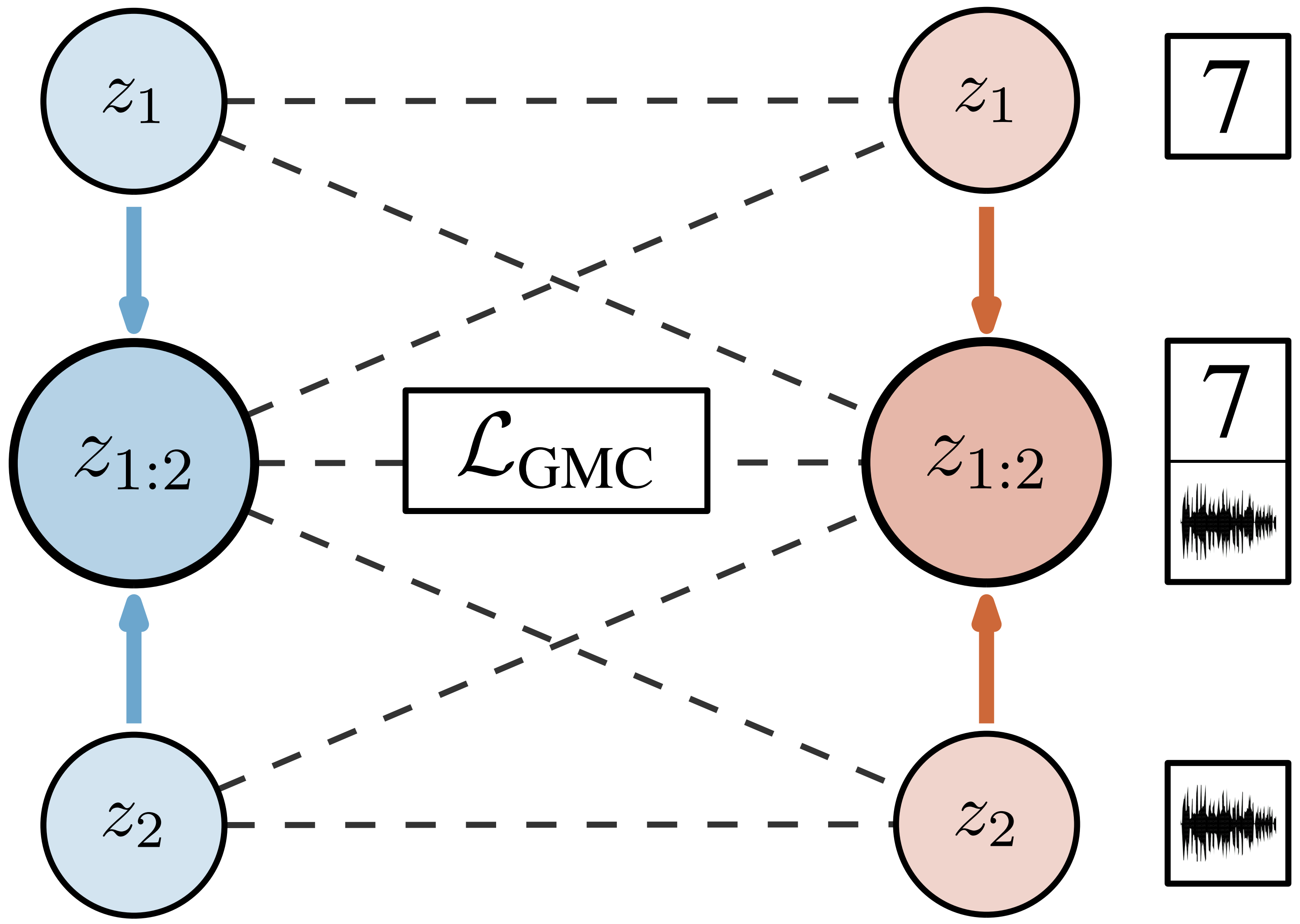


Shared network that outputs *latent* representations $\{z_{1:M}, z_1, \ldots, z_M\} \in \mathbb{R}^s$;

GMC: Method

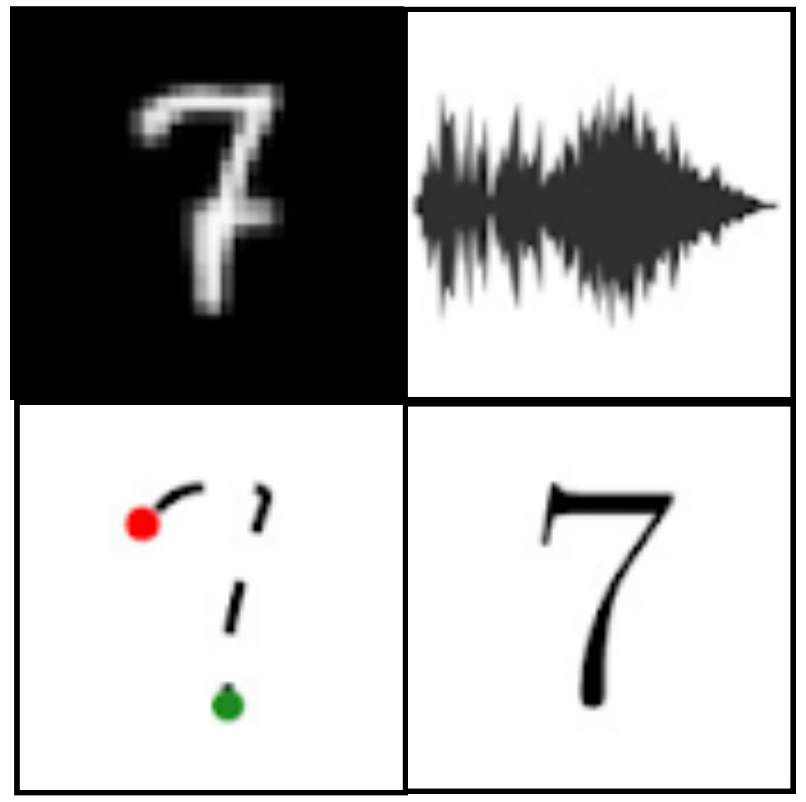


Loss Function



Multimodal contrastive loss to promote the *geometric alignment* of the complete and modality-specific representations.

Evaluation

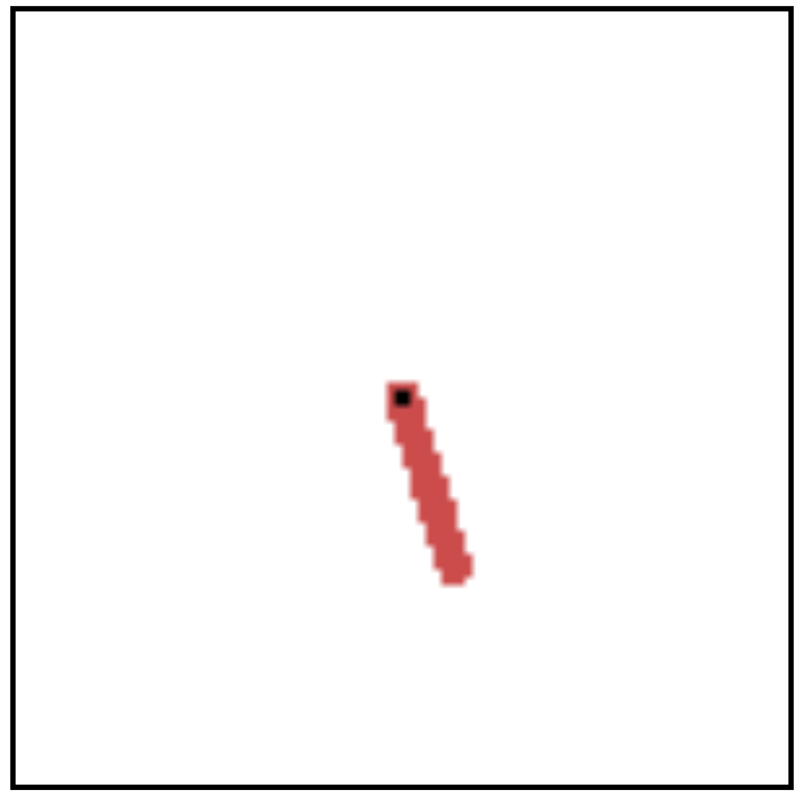


Unsupervised Learning

Dataset: *MHD [4]*
Modalities*: 4*
Downstream*: Classification*

Supervised Learning

Dataset: *CMU-MOSEI [5]*
Modalities*: 3*
Downstream*: Classification*



Reinforcement Learning

Dataset: *Multimodal Pendulum [6]*
Modalities*: 2*
Downstream*: Control*

[4] Vasco, Miguel, et al. "Leveraging hierarchy in multimodal generative models for effective cross-modality inference." *Neural Networks* (2022)
[5] Zadeh, Amir, and Paul Pu. "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph." *ACL* (2018)
[6] Silva, Rui, et al. "Playing Games in the Dark: An Approach for Cross-Modality Transfer in Reinforcement Learning." *AAMAS* (2020)

Evaluation: Unsupervised

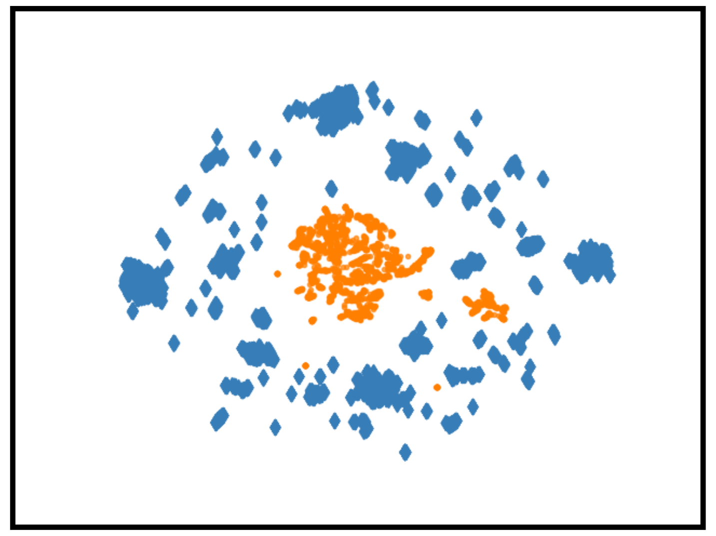
Downstream Performance: *Classification*

Table 1. Performance of different multimodal representation methods in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy (%) results averaged over 5 independent runs. Higher is better.

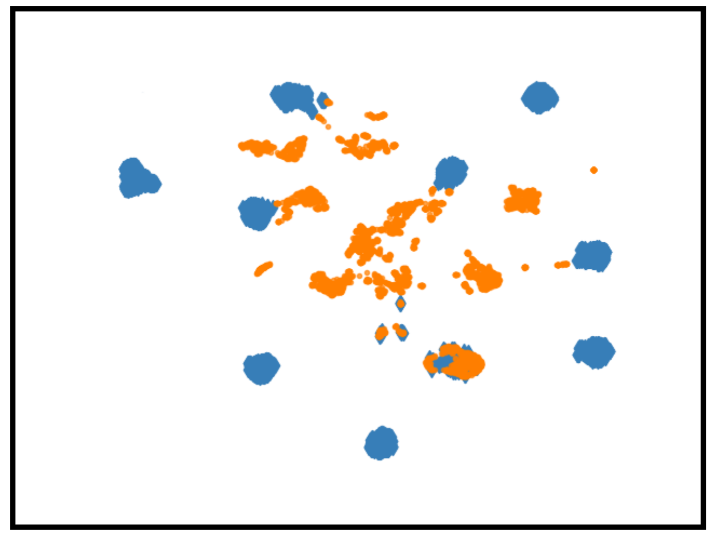
Input	MVAE[1]	MMVAE	Nexus	MUSE	MFM	**GMC (Ours)**
Complete ($x_{1:4}$)	100.0 ± 0.00	99.81 ± 0.21	99.98 ± 0.05	$99.99 \pm 4e{-5}$	100.0 ± 0.00	100.0 ± 0.00
Image (x_1)	77.94 ± 3.16	94.63 ± 2.61	95.89 ± 0.34	79.37 ± 2.75	34.66 ± 6.48	$\mathbf{99.75 \pm 0.03}$
Sound (x_2)	61.75 ± 4.59	69.43 ± 26.43	39.07 ± 5.82	41.39 ± 0.18	10.07 ± 0.20	$\mathbf{93.04 \pm 0.45}$
Trajectory (x_3)	10.03 ± 0.06	95.33 ± 2.56	98.55 ± 0.34	89.49 ± 2.44	25.61 ± 5.41	$\mathbf{99.96 \pm 0.02}$
Label (x_4)	100.0 ± 0.00	87.99 ± 7.49	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00	100.0 ± 0.00

Evaluation: Unsupervised

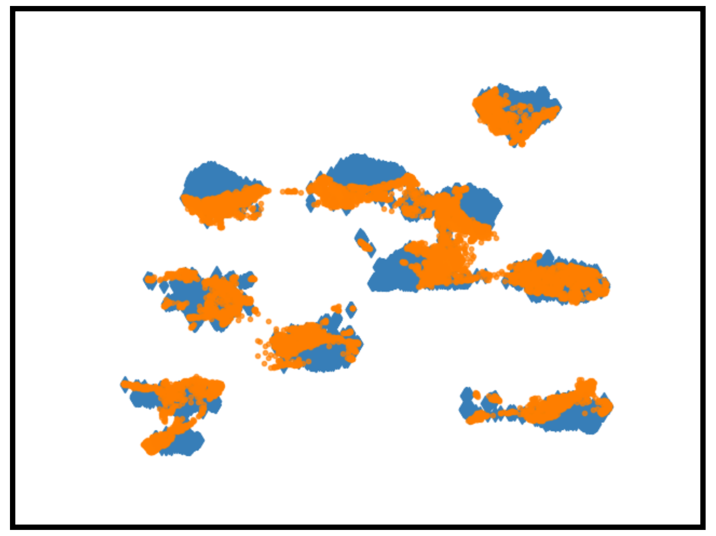
Geometric Alignment: *UMAP [7]* ● Multimodal Observations ● Image Observations



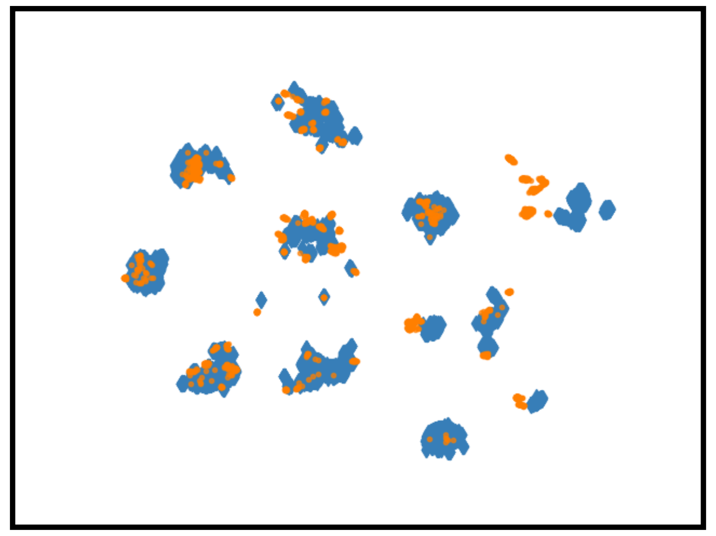
MFM
(Tsai et al., 2019)



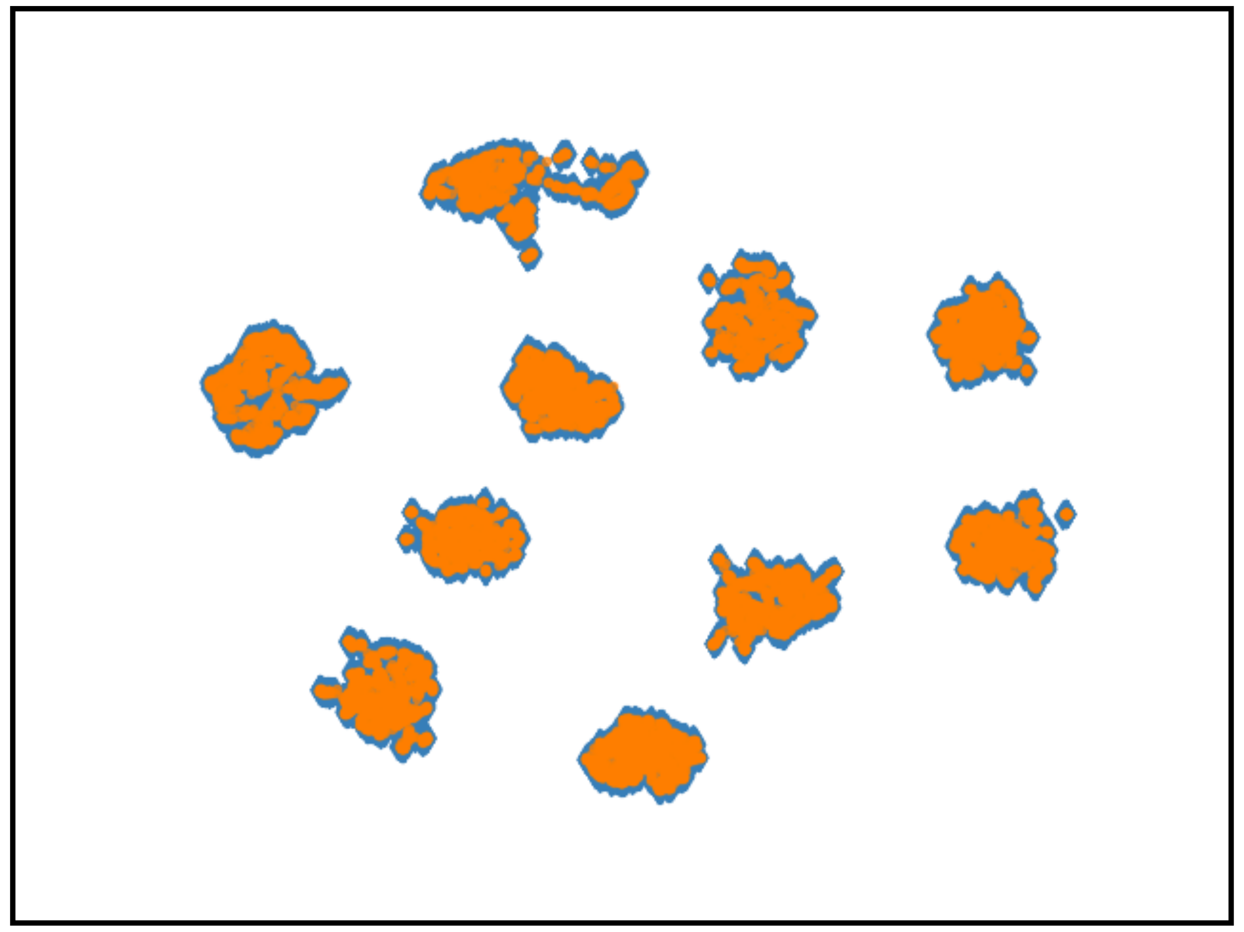
MVAE
(Wu & Goodman, 2019)



MUSE
(Vasco et al., 2022)



MMVAE
(Shi et al., 2022)



GMC *(Ours)*

[7] McInnes, Leland, et al. "UMAP: Uniform Manifold Approximation and Projection." *Journal of Open Source Software* (2018)

Evaluation: Unsupervised

Geometric Alignment: *DCA [8]*

Table 2. DCA score of the models in the MHD dataset, evaluating the geometric alignment of complete representations $z_{1:4}$ and modality-specific ones $\{z_1, \ldots, z_4\}$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	MVAE[1]	MMVAE	Nexus	MUSE	MFM	GMC (Ours)
Complete ($z_{1:4}$)	Image (z_1)	0.01 ± 0.01	0.21 ± 0.29	0.00 ± 0.00	0.54 ± 0.44	0.00 ± 0.00	$\mathbf{0.96 \pm 0.02}$
Complete ($z_{1:4}$)	Sound (z_2)	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.87 \pm 0.16}$
Complete ($z_{1:4}$)	Trajectory (z_3)	0.00 ± 0.00	0.01 ± 0.01	0.08 ± 0.02	0.00 ± 0.00	0.00 ± 0.00	$\mathbf{0.86 \pm 0.05}$
Complete ($z_{1:4}$)	Label (z_4)	0.99 ± 0.01	0.74 ± 0.22	0.43 ± 0.05	0.93 ± 0.05	0.85 ± 0.06	$\mathbf{1.00 \pm 0.00}$

[8] Poklukar, Petra, et al. "Delaunay Component Analysis for Evaluation of Data Representations." *ICLR* (2022)

Evaluation: Supervised

Downstream Performance: *Classification*

Table 4. Performance of different multimodal representation methods in the CMU-MOSEI dataset, in a classification task under complete and partial observations. Results averaged over 5 independent runs. Arrows indicate the direction of improvement.

Metric	Baseline	GMC (Ours)
MAE (↓)	0.643 ± 0.019	$\mathbf{0.634 \pm 0.008}$
Cor (↑)	$\mathbf{0.664 \pm 0.004}$	0.653 ± 0.004
F1 (↑)	$\mathbf{0.809 \pm 0.003}$	0.798 ± 0.008
Acc (%, ↑)	$\mathbf{80.75 \pm 00.28}$	79.73 ± 00.69

(a) Complete Observations ($x_{1:3}$)

Metric	Baseline	**GMC (Ours)**
MAE (↓)	0.805 ± 0.028	$\mathbf{0.712 \pm 0.015}$
Cor (↑)	0.427 ± 0.061	$\mathbf{0.590 \pm 0.013}$
F1 (↑)	0.713 ± 0.086	$\mathbf{0.779 \pm 0.005}$
Acc (%, ↑)	66.53 ± 09.86	$\mathbf{77.85 \pm 00.36}$

(b) Text Observations (x_1)

Metric	Baseline	**GMC (Ours)**
MAE (↓)	0.873 ± 0.065	$\mathbf{0.837 \pm 0.008}$
Cor (↑)	0.090 ± 0.062	$\mathbf{0.256 \pm 0.007}$
F1 (↑)	0.622 ± 0.122	$\mathbf{0.676 \pm 0.015}$
Acc (%, ↑)	53.17 ± 09.47	$\mathbf{65.59 \pm 00.62}$

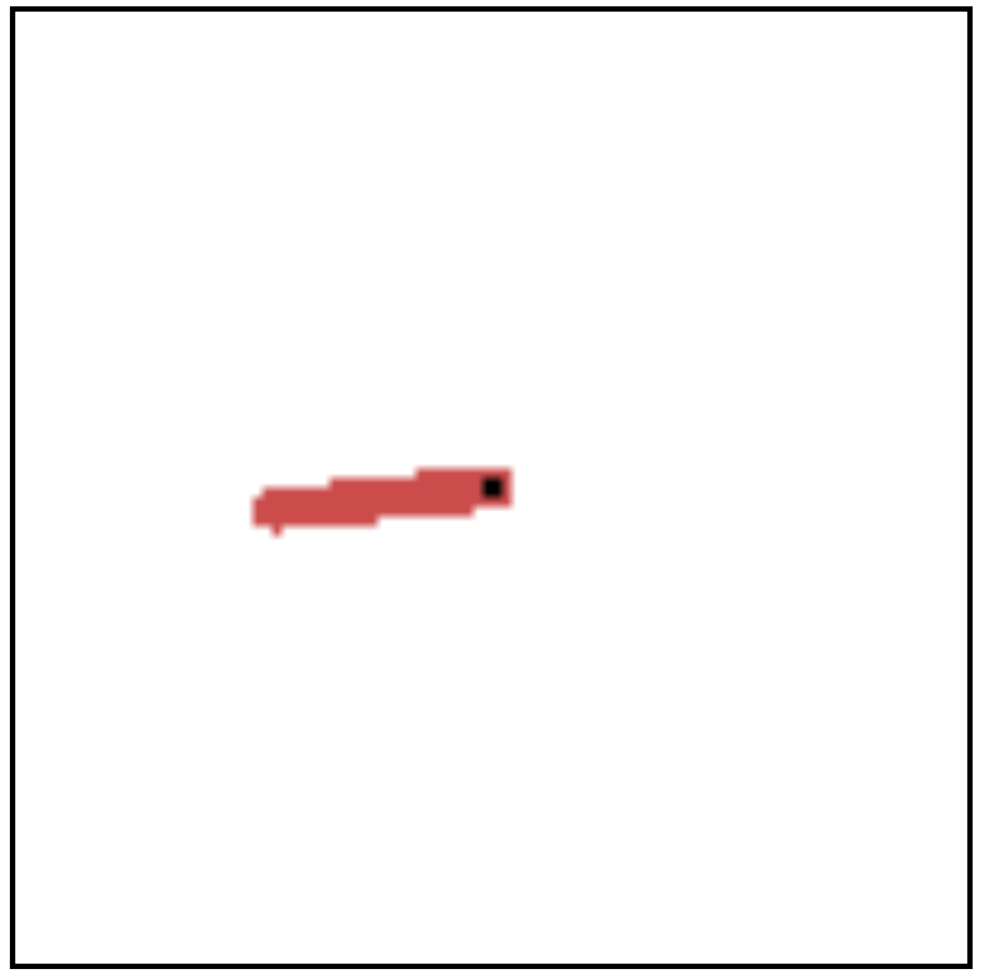
(c) Audio Observations (x_2)

Metric	Baseline	**GMC (Ours)**
MAE (↓)	1.025 ± 0.164	$\mathbf{0.845 \pm 0.010}$
Cor (↑)	0.110 ± 0.060	$\mathbf{0.278 \pm 0.011}$
F1 (↑)	0.574 ± 0.095	$\mathbf{0.655 \pm 0.003}$
Acc (%, ↑)	44.33 ± 09.40	$\mathbf{65.02 \pm 00.28}$

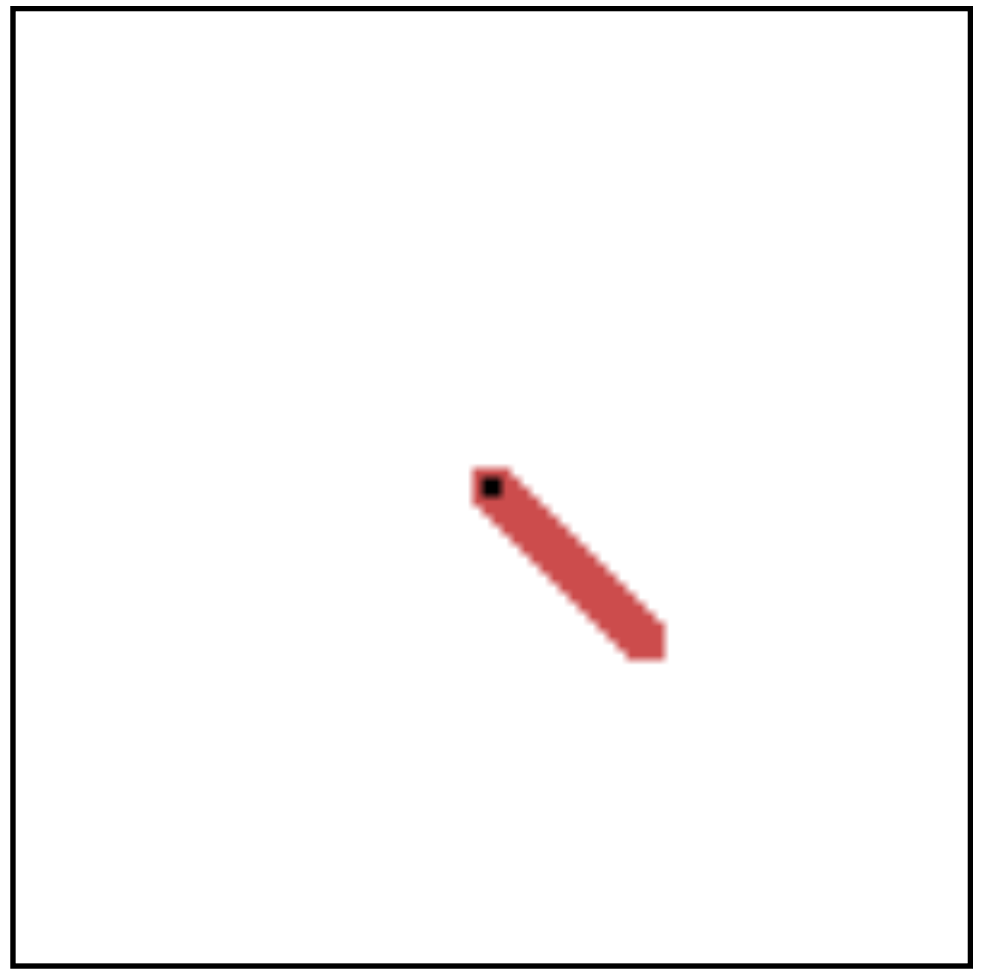
(d) Video Observations (x_3)

Evaluation: RL

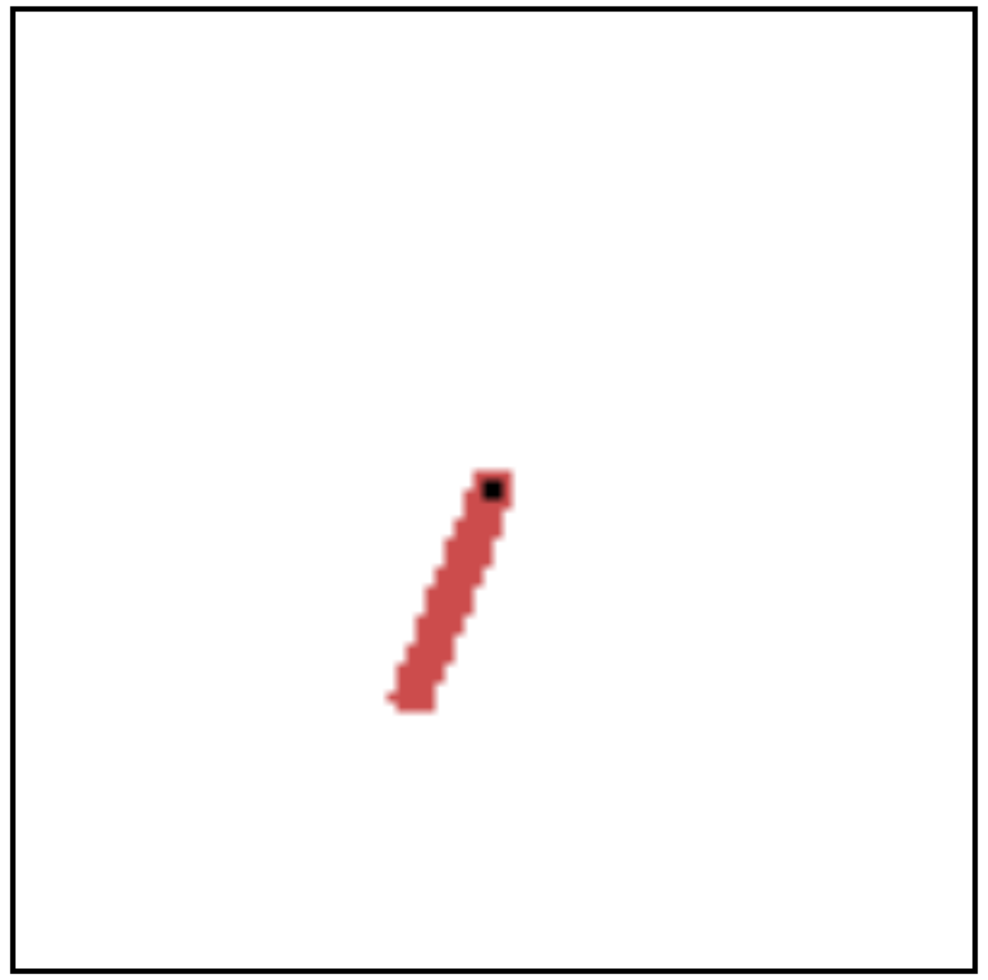
Downstream Performance: *Acting only with sound observations*



MVAE
(Wu & Goodman, 2018)



MUSE
(Vasco et al., 2022)



GMC
(Ours)

Geometric Multimodal Contrastive Representation Learning

Petra Poklukar, Miguel Vasco*, Hang Yin, Francisco S. Melo, Ana Paiva, Danica Kragic*

Read **Paper**

Get the **Code**