Geometric Multimodal Contrastive Representation Learning

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

Abstract

Learning representations of multimodal data that are both informative and robust to missing modalities at test time remains a challenging problem due to the inherent heterogeneity of data obtained from different channels. To address it, we present a novel Geometric Multimodal Contrastive (GMC) representation learning method consisting of two main components: *i)* a two-level architecture consisting of *modality-specific* base encoders, allowing to process an arbitrary number of modalities to an intermediate representation of fixed dimensionality, and a *shared* projection head, mapping the intermediate representations to a latent representation space; *ii)* a multimodal contrastive loss function that encourages the geometric alignment of the learned representations. We experimentally demonstrate that GMC representations are semantically rich and achieve state-of-the-art performance with missing modality information on three different learning problems including prediction and reinforcement learning tasks.

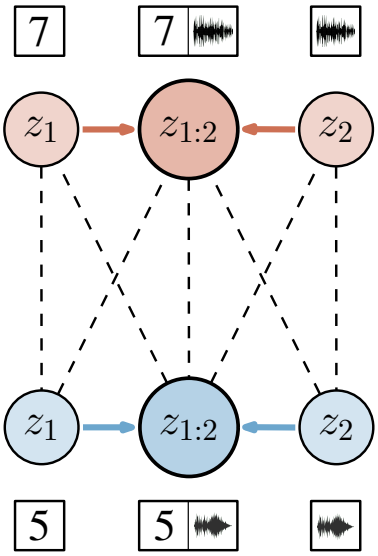


Figure 1. We propose the *Geometric Multimodal Contrastive* (GMC) framework to learn representations of multimodal data by *aligning* the corresponding modality-specific (z_1 or z_2) and complete ($z_{1:2}$) representations (solid arrows, blue circles) and *contrasting* with different modality-specific and complete pairs (dashed lines, red circles).

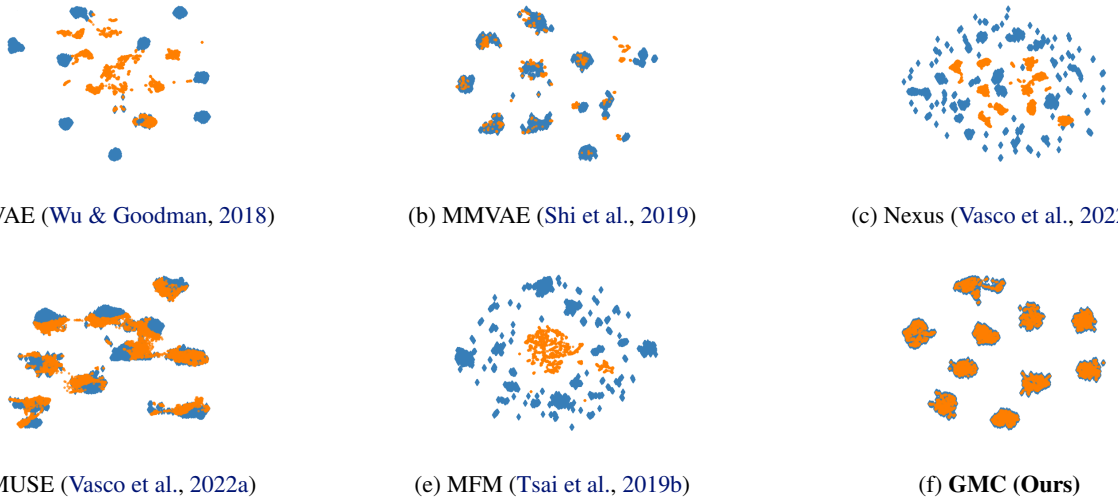
1. Introduction

Information regarding objects or environments in the world can be recorded in the form of signals of different nature. These different *modality* signals can be for instance images, videos, sounds or text, and represent the same underlying phenomena. Naturally, the performance of machine learning models can be enhanced by leveraging the redundant and complementary information provided by multiple modalities (Baltrušaitis et al., 2018). In particular, exploiting such multimodal information has been shown to be successful in tasks such as classification (Tsai et al., 2019b;a), generation (Wu & Goodman, 2018; Shi et al., 2019) and control (Silva et al., 2020; Vasco et al., 2022a).

The advances of many of these methods can be attributed to the efficient learning of multimodal data representations, which reduces the inherent complexity of raw multimodal data and enables the extraction of the underlying semantic correlations among the different modalities (Baltrušaitis et al., 2018; Guo et al., 2019). Generally, good representations of multimodal data *i) capture the semantics* from individual modalities necessary for performing a given downstream task. Additionally, in scenarios such as real-world classification and control, it is essential that the obtained representations are *ii) robust to missing modality* information during execution (Meo & Lanillos, 2021; Tremblay et al., 2021; Zambelli et al., 2020). In order to fulfill *i)* and *ii)*, the unique characteristics of each modality need to be processed accordingly and efficiently combined, which remains a challenging problem known as the *heterogeneity gap* in multimodal representation learning (Guo et al., 2019).

[*]Equal contribution [1]KTH Royal Institute of Technology, Stockholm, Sweden [2]INESC-ID & Instituto Superior Técnico, University of Lisbon, Portugal. Correspondence to: Miguel Vasco <miguel.vasco@tecnico.ulisboa.pt>, Petra Poklukar <poklukar@kth.se>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.



(a) MVAE (Wu & Goodman, 2018) (b) MMVAE (Shi et al., 2019) (c) Nexus (Vasco et al., 2022b)

(d) MUSE (Vasco et al., 2022a) (e) MFM (Tsai et al., 2019b) (f) **GMC (Ours)**

Figure 2. UMAP visualization of complete representations $z_{1:4}$ (blue) and image representations z_1 (orange) in a latent space $z \in \mathbb{R}^{64}$ obtained from several state-of-the-art multimodal representation learning models on the MHD dataset considered in Section 5.1. Only GMC is able to learn *modality-specific* and *complete* representations that are geometrically aligned. More visualizations in Appendix F.

An intuitive idea to mitigate the heterogeneity gap is to project heterogeneous data into a shared representation space such that the representations of complete observations capture the semantic content shared across all modalities. In this regard, two directions have shown promise, namely, generation-based methods commonly extending the Variational Autoencoder (VAE) framework (Kingma & Welling, 2014) to multimodal data such as MVAE (Wu & Goodman, 2018) and MMVAE (Shi et al., 2019), as well as methods relying on the fusion of modality-specific representations such as MFM (Tsai et al., 2019b) and the Multimodal Transformer (Tsai et al., 2019a). Fusion based methods by construction fulfill objective *i)* but typically do not provide a mechanism to cope with missing modalities. While this is better accounted for in the generation based methods, these approaches often struggle to align complete and modality-specific representations due to the demanding reconstruction objective. We thoroughly discuss the geometric misalignment of these methods in Section 2.

In this work, we learn *geometrically aligned* multimodal data representations that provide robust performance in downstream tasks under missing modalities at test time. To this end, we present the *Geometric Multimodal Contrastive (GMC)* representation learning framework. Inspired by the recently proposed Normalized Temperature-scaled Cross Entropy (NT-XEnt) loss in visual contrastive representation learning (Chen et al., 2020), we contribute a novel multimodal contrastive loss that explicitly aligns modality-specific representations with the representations obtained from the corresponding complete observation, as depicted in Figure 1. GMC assumes a two-level neural-network model architecture consisting of a collection of *modality-specific* base encoders, processing modality data into an intermediate representation of a fixed dimensionality, and a *shared* projection head, mapping the intermediate representations into a latent representation space where the contrastive learning objective is applied. It can be scaled to an arbitrary number of modalities, and provides semantically rich representations that are robust to missing modality information. Furthermore, as shown in our experiments, GMC is general as it can be integrated into existing models and applied to a variety of challenging problems, such as learning representations in an unsupervised manner (Section 5.1), for prediction tasks using a weak supervision signal (Section 5.2) or downstream reinforcement learning tasks (Section 5.3). We show that GMC is able to achieve state-of-the-art performance with missing modality information compared to existing models.

2. The Problem of Geometric Misalignment in Multimodal Representation Learning

We consider scenarios where information is provided in the form of a dataset X of N tuples, i.e., $X = \{x^i_{1:M} = (x^i_1, \dots, x^i_M)\}_{i=1}^N$, where each tuple $x_{1:M} = (x_1, \dots, x_M)$ represents observations provided by M different modalities. We refer to the tuples $x_{1:M}$ consisting of all M modalities as *complete* observations and to the single observations x_m as *modality-specific*. The goal is to learn complete representations $z_{1:M}$ of $x_{1:M}$ and modality-specific representations $\{z_1, \dots, z_M\}$ of $\{x_1, \dots, x_M\}$ that are:

i) informative, i.e., both $z_{1:M}$ and any of

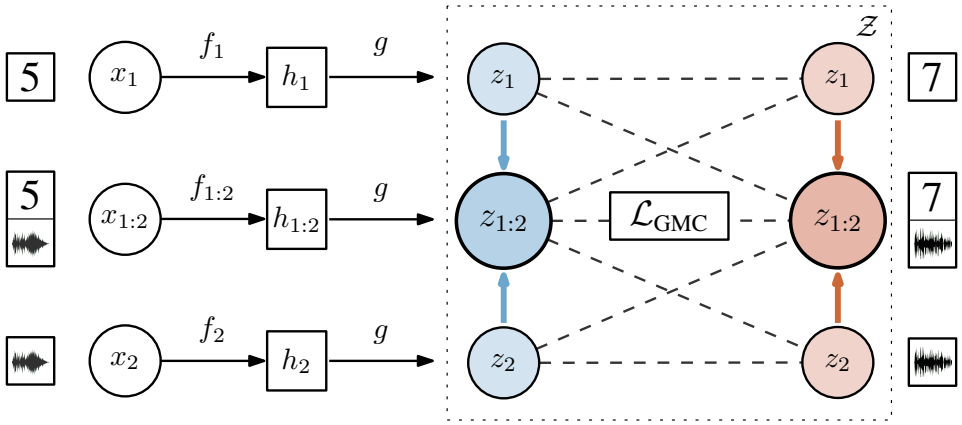


Figure 3. The *Geometric Multimodal Contrastive* (GMC) framework instantiated in scenarios with two modalities ($M = 2$): *modality-specific* base networks $f(\cdot) = \{f_{1:2}(\cdot)\} \cup \{f_1(\cdot), f_2(\cdot)\}$ encode common-dimensionality intermediate representations h that are projected using a *shared* projection head $g(\cdot)$ to a common representation space $\mathcal{Z}$, in which we apply a novel multimodal contrastive loss $\mathcal{L}_{\text{GMC}}$, detailed in Eq. (2), that *aligns* corresponding modality-specific $\{z_1, z_2\}$ and complete $z_{1:2}$ representations (coloured arrows) and *contrasts* with representations from different observations (dashed lines).

$z_m \in \{z_1, \ldots, z_M\}$ contains relevant semantic information for some downstream task, and thus,

ii) robust to missing modalities during test time, i.e., the success of a subsequent downstream task is independent of whether the provided input is the complete representation $z_{1:M}$ or any of the modality-specific representations $z_m \in \{z_1, \ldots, z_M\}$.

Prior work has demonstrated success in using complete representations $z_{1:M}$ in a diverse set of applications, such as image generation (Wu & Goodman, 2018; Shi et al., 2019) and control of Atari games (Silva et al., 2020; Vasco et al., 2022a). Intuitively, if complete representations $z_{1:M}$ are sufficient to perform a downstream task then learning modality-specific representations that are geometrically aligned with $z_{1:M}$ in the same representation space should ensure that z_m contain necessary information to perform the task even when $z_{1:M}$ cannot be provided. Therefore, in Section 5 we study the geometric alignment of $z_{1:M}$ and each z_m on several multimodal datasets and state-of-the-art multimodal representation learning models. In Figure 2, we visualize an example of encodings of $z_{1:M}$ (in blue) and z_m corresponding to the image modality (in orange) where we see that the existing approaches produce geometrically misaligned representations. As we empirically show in Section 5, this misalignment is consistent across different learning scenarios and datasets, and can lead to a poor performance on downstream tasks.

To fulfill *i)* and *ii)*, we propose a novel approach that builds upon the simple idea of geometrically aligning modality-specific representations z_m with the corresponding complete representations $z_{1:M}$ in a latent representation space, framing it as a contrastive learning problem.

3. Geometric Multimodal Contrastive Learning

We present the *Geometric Multimodal Contrastive* (GMC) framework, visualized in Figure 3, consisting of three main components:

- A collection of neural network *base* encoders $f(\cdot) = \{f_{1:M}(\cdot)\} \cup \{f_1(\cdot), \ldots, f_M(\cdot)\}$, where $f_{1:M}(\cdot)$ and $f_m(\cdot)$ take as input the complete $x_{1:M}$ and modality-specific observations x_m, respectively, and output intermediate d-dimensional representations $\{h_{1:M}, h_1, \ldots, h_M\} \in \mathbb{R}^d$;

- A neural network *shared* projection head $g(\cdot)$ that maps the intermediate representations given by the base encoders $f(\cdot)$ to the latent representations $\{z_{1:M}, z_1, \ldots, z_M\} \in \mathbb{R}^s$ over which we apply the contrastive term. The projection head $g(\cdot)$ enables to encode the intermediate representations in a shared representation space $\mathcal{Z}$ while preserving modality-specific semantics;

- A multimodal contrastive NT-Xent loss function ($\mathcal{L}_{\text{GMC}}$), that is inspired by the recently proposed SimCLR framework (Chen et al., 2020) and encourages the geometric alignment of z_m and $z_{1:M}$.

We phrase the problem of geometrically aligning z_m with $z_{1:M}$ as a contrastive prediction task where the goal is to identify z_m and its corresponding complete representation $z_{1:M}$ in a given mini-batch. Let $\mathcal{B} = \{z^i_{1:M}\}_{i=1}^B \subset g(f(X))$ be a mini-batch of B complete representations. Let $\text{sim}(u, v)$ denote the cosine similarity among vectors u and

v and let $\tau \in (0, \infty)$ be the temperature hyperparameter. We denote by

$$s_{m,n}(i,j) = \exp(\text{sim}(z_m^i, z_n^j)/\tau), \tag{1}$$

the similarity between representations z_m^i and z_n^j (modality-specific or complete) corresponding to the ith and jth samples from the mini-batch $\mathcal{B}$. For a given modality m, we define positive pairs as $(z_m^i, z_{1:M}^i)$ for $i = 1, \dots, B$ and treat the remaining pairs as negative ones. In particular, we denote by

$$\Omega_m(i) = \sum_{i \neq j} (s_{m,1:M}(i,j) + s_{m,m}(i,j) + s_{1:M,1:M}(i,j)),$$

the sum of similarities among negative pairs that correspond to the positive pair $(z_m^i, z_{1:M}^i)$, and define the contrastive loss for the same pair of samples as

$$l_m(i) = -\log \frac{s_{m,1:M}(i,i)}{\Omega_m(i)}.$$

Lastly, we combine the loss terms for each modality $m = 1, \dots, M$ and obtain the final training loss

$$\mathcal{L}_{\text{GMC}}(\mathcal{B}) = \sum_{m=1}^{M} \sum_{i=1}^{B} l_m(i). \tag{2}$$

As we only contrast single modality-specific representations to the complete ones, $\mathcal{L}_{\text{GMC}}$ scales linearly to an arbitrary number of modalities. In Section 5, we show that $\mathcal{L}_{\text{GMC}}$ can be added as an additional term to existing frameworks to improve their robustness to missing modalities. Moreover, we experimentally demonstrate that the architectures of the base encoders and shared projection head can be flexibly adjusted depending on the task.

4. Related Work

Learning multimodal representations suitable for downstream tasks has been extensively addressed in literature (Baltrušaitis et al., 2018; Guo et al., 2019). In this work, we focus on the problem of aligning modality-specific representations in a (shared) latent space emerging from the heterogeneity gap between different data sources. Prior work promoting such alignment can be separated into two groups: generation-based methods adjusting Variational Autoencoder (VAE) (Kingma & Welling, 2014) frameworks that considers a prior distribution over the shared latent space, and fusion-based methods that merge modality-specific representations into a shared representation.

Generation-based methods Associative VAE (AVAE) (Yin et al., 2017) and Joint Multimodal VAE (JMVAE) (Suzuki et al., 2016) explicitly enforce the alignment of modality-specific representations by minimizing the Kullback–Leibler divergence between their distributions. However, these models are not easily scalable to large number of modalities due to the combinatorial increase of inference networks required to account for all subsets of modalities. In contrast, GMC scales linearly with the number of modalities as it separately contrasts individual modality-specific representations to the complete ones.

Other multimodal VAE models promote the approximation of modality-specific representations through dedicated training schemes. MVAE (Wu & Goodman, 2018) uses subsampling to learn a joint-modality representation obtained from a Product-of-Experts (PoE) inference network. This solution is prone to learning overconfident experts, hindering both the alignment of the modality-specific representations and the performance of downstream tasks under incomplete information (Shi et al., 2019). Mixture-of-Experts MVAE (MMVAE) (Shi et al., 2019) instead employs a doubly reparameterized gradient estimator which is computationally expensive compared to the lower-bound objective of traditional multimodal VAEs because of its Monte-Carlo-based training scheme. GMC, on the other hand, presents an efficient training scheme without suffering from modality-specific biases.

Recently, hierarchical multimodal VAEs have been proposed to facilitate the learning of aligned multimodal representations such as Nexus (Vasco et al., 2022b) and Multimodal Sensing (MUSE) (Vasco et al., 2022a). Nexus considers a two-level hierarchy of modality-specific and multimodal representation spaces employing a dropout-based training scheme. The average aggregator solution employed to merge multimodal information lacks expressiveness which hinders the performance of the model on downstream tasks. To address this issue, MUSE introduces a PoE solution that merges lower-level modality-specific information to encode a high-level multimodal representation, and a dedicated training scheme to counter the overconfident expert issue. In contrast to both solutions, GMC is computationally efficient without requiring hierarchy.

Fusion-based methods Other class of methods approach the alignment of modality-specific representations through complex fusion mechanisms (Liang et al., 2021). The Multimodal Factorized model (MFM) (Tsai et al., 2019b) proposes the factorization of a multimodal representation into distinct multimodal discriminative factors and modality-specific generative factors, which are subsequently fused for downstream tasks. More recently, the Multimodal Transformer model (Tsai et al., 2019a) has shown remarkable classification performance in multimodal time-series datasets, employing a directional pairwise cross-modal attention mechanism to learn a rich representation of heterogeneous data streams without requiring their explicit time-alignment. In contrast to both models, GMC is able to learn multimodal

representations of modalities of arbitrary nature without explicitly requiring a supervision signal (e.g. labels).

5. Experiments

We evaluate the quality of the representations learned by GMC on three different scenarios:

- An *unsupervised learning* problem, where we learn multimodal representations on the Multimodal Handwritten Digits (MHD) dataset (Vasco et al., 2022b). We showcase the geometric alignment of representations and demonstrate the superior performance of GMC compared to the baselines on a downstream classification task with missing modalities (Section 5.1);
- A *supervised learning* problem, where we demonstrate the flexibility of GMC by integrating it into state-of-the-art approaches to provide robustness to missing modalities in challenging classification scenarios (Section 5.2);
- A *reinforcement learning* (RL) task, where we show that GMC produces general representations that can be applied to solve downstream control tasks and demonstrate state-of-the-art performance in actuation with missing modality information (Section 5.3).

In each corresponding section, we describe the dataset, baselines, evaluation and training setup used. We report all model architectures and training hyperparameters in Appendix D and E. All results are averaged over 5 different randomly-seeded runs except for the RL experiments where we consider 10 runs. Our code is available on GitHub[2].

Evaluation of geometric alignment To evaluate the geometric alignment of representations, we use a recently proposed *Delaunay Component Analysis* (DCA) (Poklukar et al., 2022) method designed for general evaluation of representations. DCA is based on the idea of comparing geometric and topological properties of an evaluation set of representations E with the reference set R, acting as an approximation of the true underlying manifold. The set E is considered to be well aligned with R if its global and local structure resembles well the one captured by R, i.e., the manifolds described by the two sets have similar number, structure and size of connected components.

DCA approximates the manifolds described by R and E with a Delaunay neighbourhood graph and derives several scores reflecting their alignment. We consider three of them: *network quality* $q \in [0, 1]$ which measures the overall geometric alignment of R and E in the connected components, as well as *precision* $\mathcal{P} \in [0, 1]$ and *recall* $\mathcal{R} \in [0, 1]$ which measure the proportion of points from E and R, respectively, that are contained in geometrically well-aligned components. To account for all three normalized scores, we report the harmonic mean defined as $3/(1/\mathcal{P} + 1/\mathcal{R} + 1/q)$ when all $\mathcal{P}, \mathcal{R}, q > 0$ and 0 otherwise. In all experiments, we compute DCA using complete representations $z_{1:M}$ as the reference set R and modality-specific z_m as the evaluation set E, both obtained from testing observations. A detailed description of the method and definition of the scores is found in Appendix A.

5.1. Experiment 1: Unsupervised Learning

Datasets The MHD dataset is comprised of images (x_1), sounds (x_2), motion trajectories (x_3) and label information (x_4) related to handwriting digits. The authors collected $60,000$ 28×28 greyscale images per class as well as normalized 200-dimensional representations of trajectories and 128×32-dimensional representations of audio. The dataset is split into $50,000$ training and $10,000$ testing samples.

Models We consider several generation-based and fusion-based state-of-the-art multimodal representation methods: MVAE, MMVAE, Nexus, MUSE and MFM (see Section 4 for a detailed description). For a fair comparison, when possible, we employ the same encoder architectures and latent space dimensionality across all baseline models, described in Appendix D. For GMC, we employ the same modality-specific base encoders $f_m(\cdot)$ as the baselines with an additional base encoder $f_{1:4}(\cdot)$ taking complete observations as input. The shared projection head $g(\cdot)$ comprises of 3 fully-connected layers. We set the temperature $\tau = 0.1$ and consider 64-dimensional intermediate and shared representation spaces, i.e., $h \in \mathbb{R}^{64}, z \in \mathbb{R}^{64}$. We train all the models for 100 epochs using a learning rate of 10^{-3}, employing the training schemes and hyperparameters suggested by the authors (when available).

Evaluation We follow the established evaluation in the literature using classification as a downstream task (Shi et al., 2019) and train a 10-class classifier neural network on complete representations $z_{1:M} = g\left(f_{1:M}(x_{1:M})\right)$ from the training split (see Appendix D for the exact architecture). The classifier is trained for 50 epochs using a learning rate of $1\mathrm{e}{-3}$. We report the testing accuracy obtained when the classifier is provided with both complete $z_{1:4}$ and modality-specific representations z_m as inputs.

Classification results The classification results are shown in Table 1. While all the models attain perfect accuracy on $x_{1:4}$ and x_4, we observe that GMC is the only model that successfully performs the task when given only x_1, x_2 or x_3 as input, significantly outperforming the baselines.

Geometric alignment To validate that the superior perfor-

[1]Results averaged over 3 randomly-seeded runs due to divergence during MVAE training in the remaining seeds.

[2]https://github.com/miguelsvasco/gmc

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

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	$\mathbf{1.00 \pm 0.00}$

Table 3. Number of parameters (in millions) of the representation models employed in the Multimodal Handwritten Digits dataset.

MVAE	MMVAE	Nexus	MUSE	MFM	**GMC (Ours)**
9.3	9.0	12.9	9.9	9.4	**2.9**

mance of GMC originates from a better geometric alignment of representations, we evaluate the testing representations obtained from all the models using DCA. For each modality m, we compared the alignment of the evaluation set $E = \{z_m\}$ and the reference set $R = \{z_{1:4}\}$. The obtained DCA scores are shown in Table 2 where we see that GMC outperforms all the considered baselines. For some cases, we observe the obtained representations are completely misaligned yielding $\mathcal{P} = \mathcal{R} = q = 0$. While some of the baselines are to some extend able to align z_1 and/or z_4 to $z_{1:4}$, GMC is the only method that is able to align even the sound and trajectory representations, z_2 and z_3, resulting in a superior classification performance.

We additionally validate the geometric alignment by visualizing 2-dimensional UMAP projections (McInnes et al., 2018) of the representations z. In Figure 2 we show projections of $z_{1:4}$ and image representations z_1 obtained using the considered models. We clearly see that GMC not only correctly aligns $z_{1:4}$ and z_1 but also separates the representations in 10 clusters. Moreover, we can see that among the baselines only MMVAE and MUSE somewhat align the representations which is on par with the quantitative results reported in Table 2. For MVAE, Nexus and MFM, Figure 2 visually supports the obtained DCA score 0. Note that points marked as outliers by DCA are omited from the visualization. We provide similar visualizations of other modalities in Appendix F.

Model Complexity In Table 3 we present the number of parameters required by the multimodal representation models employed in this task. The results show that GMC requires significantly fewer parameters than the smallest baseline model – 68% fewer parameters than MMVAE.

5.2. Experiment 2: Supervised Learning

In this section, we evaluate the flexibility of GMC by adjusting both the architecture of the model and training procedure to receive an additional supervision signal during training to guide the learning of complete representations. We demonstrate how GMC can be integrated into existing approaches to provide additional robustness to missing modalities with minimal computational cost.

Datasets We employ the CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Bagher Zadeh et al., 2018), two popular datasets for sentiment analysis and emotion recognition with challenging temporal dynamics. Both datasets consist of textual (x_1), sound (x_2) and visual (x_3) modalities extracted from videos. CMU-MOSI consists of 2199 short monologue videos clips of subjects expressing opinions about various topics. CMU-MOSEI is an extension of CMU-MOSI dataset containing 23453 YouTube video clips of subjects expressing movie reviews. In both datasets, each video clip is annotated with labels in $[-3, 3]$, where -3 and 3 indicate strong negative and strongly positive sentiment scores, respectively. We employ the temporally-aligned

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)

Table 5. DCA score of the models in the CMU-MOSEI dataset evaluating the geometric alignment of complete representations $z_{1:4}$ and modality-specific ones $\{z_1, z_2, z_3\}$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	Baseline	**GMC (Ours)**
Complete ($z_{1:3}$)	Text (z_1)	0.50 ± 0.05	$\mathbf{0.95 \pm 0.01}$
Complete ($z_{1:3}$)	Audio (z_2)	0.41 ± 0.14	$\mathbf{0.86 \pm 0.04}$
Complete ($z_{1:3}$)	Vision (z_3)	0.50 ± 0.14	$\mathbf{0.92 \pm 0.02}$

version of these datasets: CMU-MOSEI consists of 18134 and 4643 training and testing samples, respectively, and CMU-MOSI consists of 1513 and 686 training and testing samples, respectively.

Models We consider the Multimodal Transformer (Tsai et al., 2019a) which is the state-of-the-art model for classification on the CMU-MOSI and CMU-MOSEI datasets (we refer to Tsai et al. (2019a) for a detailed description of the architecture). For GMC, we employ the same architecture for the joint-modality encoder $f_{1:3}(\cdot)$ as the Multimodal Transformer but remove the last classification layers. For the modality-specific base encoders $\{f_1(\cdot), f_2(\cdot), f_3(\cdot)\}$, we employ a simple GRU layer with 30 hidden units and a fully-connected layer. The shared projection head $g(\cdot)$ is comprised of a single fully connected layer. We set $\tau = 0.3$ and consider 60-dimensional intermediate and shared representations $h, z \in \mathbb{R}^{60}$.

In addition, we employ a simple classifier consisting of 2 linear layers over the complete representations $z_{1:M}$ to provide the supervision signal to the model during training. We follow the training scheme proposed by Tsai et al. (2019a) and train all models for 40 epochs with a decaying learning rate of 10^{-3}.

Evaluation We evaluate the performance of representation learning models in sentiment analysis classification with missing modality information. We consider the same metrics as in Tsai et al. (2019b;a) and report binary accuracy (Acc), mean absolute error (MAE), correlation (Cor) and F1 score (F1) of the predictions obtain on the test dataset. In Appendix C we present similar results on the CMU-MOSI dataset.

Results The results obtained on CMU-MOSEI are reported in Table 4. When using the complete observations $x_{1:3}$ as inputs, GMC achieves competitive performance with the baseline model indicating that the additional contrastive loss does not deteriorate the model's capabilities (Table 4a). However, GMC significantly improves the robustness of the model to the missing modalities as seen in Tables 4b, 4c and 4d where we use only individual modalities as inputs. While GMC consistently outperforms the baseline in all metrics, we observe the largest improvement on the F1 score and binary accuracy (Acc) where the baseline often performs worse than random. As before, we additionally evaluate the geometric alignment of the modality-specific representations z_m (comprising the set E) and complete representations $z_{1:3}$ (comprising the set R). The resulting DCA score, reported in Table 5, supports the results shown in Table 4 and verifies that GMC significantly improves the geometric alignment compared to the baseline. Furthermore, GMC incurs in a small computational cost (with 1.4 million parameters), requiring only 300K extra parameters in comparison with the baseline (with 1.1 million parameters).

Table 6. Performance after zero-shot policy transfer in the multimodal Pendulum task. At test time, the agent is provided with either image (x_1), sound (x_2), or complete ($x_{1:2}$) observations. Total reward averaged over 100 episodes and 10 randomly seeded runs. Higher is better.

Observation	MVAE + DDPG	MUSE + DDPG	**GMC + DDPG (Ours)**
Complete ($x_{1:2}$)	-1.114 ± 0.110	-1.005 ± 0.117	$\mathbf{-0.935 \pm 0.057}$
Image (x_1)	-1.116 ± 0.121	-4.752 ± 0.994	$\mathbf{-0.940 \pm 0.056}$
Sound (x_2)	-6.642 ± 0.106	-3.459 ± 0.519	$\mathbf{-0.956 \pm 0.075}$

Table 7. DCA score of the models in the multimodal Pendulum task evaluating the geometric alignment of complete representations $z_{1:2}$ and modality-specific ones $\{z_1, z_2\}$ used as R and E inputs in DCA, respectively. Results averaged over 10 independent runs. Higher is better.

R	E	MVAE + DDPG	MUSE + DDPG[2]	**GMC + DDPG (Ours)**
Complete ($z_{1:2}$)	Image (z_1)	0.79 ± 0.01	0.20 ± 0.09	$\mathbf{0.87 \pm 0.01}$
Complete ($z_{1:2}$)	Sound (z_2)	0.00 ± 0.00	0.01 ± 0.01	$\mathbf{0.88 \pm 0.02}$

Table 8. Number of parameters (in millions) of the representation models employed in the multimodal Pendulum scenario.

MVAE	MUSE	**GMC (Ours)**
3.8	4.3	**1.9**

5.3. Experiment 3: Reinforcement Learning

In this section, we demonstrate how GMC can be employed as a representation model in the design of RL agents yielding state-of-the-art performance using missing modality information during task execution.

Scenario We consider the recently proposed multimodal inverted Pendulum task (Silva et al., 2020) which is an extension of the classical control scenario to a multimodal setting. In this task, the goal is to swing the pendulum up so it remains balanced upright. The observations of the environment include both an image (x_1) and a sound (x_2) component. The sound component is generated by the tip of the pendulum emitting a constant frequency f_0. This frequency is received by a set of S sound receivers $\{\rho_1, \dots, \rho_S\}$. At each timestep, the frequency f_i' heard by each sound receiver ρ_i is modified by the Doppler effect, modifying the frequency heard by an observer as a function of the velocity of the sound emitter. The amplitude is modified as function of the relative position of the emitter in relation to the observer following an inverse square law. To train the representation models, we employ a random policy to collect a dataset composed of 20,000 training samples and 2,000 test samples following the procedure of Silva et al. (2020).

Models We consider the MVAE (Wu & Goodman, 2018) and the MUSE (Vasco et al., 2022a) models which are two commonly used approaches for the perception of multimodal RL agents. For GMC, we employ the same modality-specific encoders $f_1(\cdot), f_2(\cdot)$ as the baselines in addition to a joint-modality encoder $f_{1:2}(\cdot)$. The shared projection head $g(\cdot)$ is comprised of 2 fully-connected layers. We use $\tau = 0.3$ and set the dimensions of intermediate and latent representations spaces to $d = 64$ and $s = 10$. We follow the two-stage agent pipeline proposed in Higgins et al. (2017) and initially train all representation models on the dataset of collected observations for 500 epochs using a learning rate of 10^{-3}. We subsequently train a Deep Deterministic Policy Gradient (DDPG) controller (Lillicrap et al., 2015) that takes as input the representations $z_{1:2}$ encoded from complete observations $x_{1:2}$ following the network architecture and training hyperparameters used by Silva et al. (2020).

Evaluation We evaluate the performance of RL agents acting under incomplete perceptions that employ the representation models to encode raw observations of the environment. During execution, the environment may provide any of the modalities $\{x_{1:2}, x_1, x_2\}$. As such, we compare the performance of the RL agents when directly using the policy learned from complete observations in scenarios with possible missing modalities without any additional training (zero-shot transfer).

Results Table 6 summarizes the total reward collected per episode for the Pendulum scenario averaged over 100 episodes and 10 randomly seeded runs[3].

The results show that only GMC is able to provide the agent with a representation model robust to partial observations allowing the agent to act under incomplete perceptual con-

[3]Results averaged over 9 randomly-seeded runs for the MUSE + DDPG method due to divergence during training in the remaining seed.

ditions with no performance loss. This is on par with the DCA scores reported in Table 7 indicating that GMC geometrically better aligns the representations compared to the baselines. Once again, as shown in Table 8, GMC can achieve such performance with 50% fewer parameters than the smallest baseline, evidence of its efficiency.

5.4. Ablation studies

We perform an ablation study on the hyperparameters of GMC using the setup from Section 5.1 on MHD dataset. In particular, we investigate: *i)* the robustness of the GMC framework when varying the temperature parameter τ; *ii)* the performance of GMC when varying dimensionalities d and s of the intermediate and latent representation spaces, respectively; and *iii)* the performance of GMC trained with a modified loss function that uses only complete observations as negative pairs. We report both classification results and DCA scores in Appendix B and observe that GMC is robust to different experimental conditions both in terms of performance and geometric alignment of representations.

6. Conclusion

We addressed the problem of learning multimodal representations that are both semantically rich and robust to missing modality information. We contributed with a novel Geometric Multimodal Contrastive (GMC) learning framework that is inspired by the visual contrastive learning methods and geometrically aligns complete and modality-specific representations in a shared latent space. We have shown that GMC is able to achieve state-of-the-art performance with missing modality information across a wide range of different learning problems while being computationally efficient (often requiring 90% fewer parameters than similar models) and straightforward to integrate with existing state-of-the-art approaches. We believe that GMC broadens the range of possible applications of contrastive learning methods to multimodal scenarios and opens many future work directions, such as investigating the effect of modality-specific augmentations or usage of inherent intermediate representations for modality-specific downstream tasks.

Acknowledgements

This work has been supported by the Knut and Alice Wallenberg Foundation, Swedish Research Council and European Research Council. This work was also partially supported by Portuguese national funds through the Portuguese Fundação para a Ciência e a Tecnologia under project UIDB/50021/2020 (INESC-ID multi annual funding) and project PTDC/CCI-COM/5060/2021. In addition, this research was partially supported by TAILOR, a project funded by EU Horizon 2020 research and innovation programme under GA No. 952215. This work was also supported by funds from Europe Research Council under project BIRD 884887. Miguel Vasco acknowledges the Fundação para a Ciência e a Tecnologia PhD grant SFRH/BD/139362/2018.

References

Bagher Zadeh, A., Liang, P. P., Poria, S., Cambria, E., and Morency, L.-P. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2):423–443, 2018.

Cao, Y.-H. and Wu, J. Rethinking self-supervised learning: Small is beautiful. *arXiv preprint arXiv:2103.13559*, 2021.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Guo, W., Wang, J., and Wang, S. Deep multimodal representation learning: A survey. *IEEE Access*, 7:63373–63394, 2019.

Higgins, I., Pal, A., Rusu, A., Matthey, L., Burgess, C., Pritzel, A., Botvinick, M., Blundell, C., and Lerchner, A. Darla: Improving zero-shot transfer in reinforcement learning. In *Procedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1480–1490, 2017.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Liang, P., Lyu, Y., Fan, X., Wu, Z., Cheng, Y., Wu, J., Chen, L., Wu, P., Lee, M., Zhu, Y., et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (Neurips)*, 2021.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

McInnes, L., Healy, J., Saul, N., and Großberger, L. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018.

Meo, C. and Lanillos, P. Multimodal vae active inference controller. In *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2693–2699. IEEE, 2021.

Poklukar, P., Varava, A., and Kragic, D. Geomca: Geometric evaluation of data representations. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 8588–8598, 2021.

Poklukar, P., Polianskii, V., Varava, A., Pokorny, F., and Kragic, D. Delaunay component analysis for evaluation of data representations. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

Shi, Y., Siddharth, N., Paige, B., and Torr, P. H. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (Neurips)*, pp. 15718–15729, 2019.

Silva, R., Vasco, M., Melo, F. S., Paiva, A., and Veloso, M. Playing games in the dark: An approach for cross-modality transfer in reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1260–1268, 2020.

Suzuki, M., Nakayama, K., and Matsuo, Y. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

Tremblay, J.-F., Manderson, T., Noca, A., Dudek, G., and Meger, D. Multimodal dynamics modeling for off-road autonomous vehicles. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1796–1802. IEEE, 2021.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6558–6569, 2019a.

Tsai, Y.-H. H., Liang, P. P., Zadeh, A., Morency, L.-P., and Salakhutdinov, R. Learning factorized multimodal representations. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019b.

Vasco, M., Yin, H., Melo, F. S., and Paiva, A. How to sense the world: Leveraging hierarchy in multimodal perception for robust reinforcement learning agents. In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 1301–1309, 2022a.

Vasco, M., Yin, H., Melo, F. S., and Paiva, A. Leveraging hierarchy in multimodal generative models for effective cross-modality inference. *Neural Networks*, 146:238–255, 2022b.

Wu, M. and Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems (Neurips)*, pp. 5580–5590, 2018.

Yin, H., Melo, F. S., Billard, A., and Paiva, A. Associate latent encodings in learning from demonstrations. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

Zambelli, M., Cully, A., and Demiris, Y. Multimodal representation models for prediction and control from partial information. *Robotics and Autonomous Systems*, 123: 103312, 2020.

A. Delaunay Component Analysis

Delaunay Component Analysis (DCA) is a recently proposed method for general evaluation of data representations (Poklukar et al., 2022). The basic idea of DCA is to compare geometric and topological properties of two sets of representations – a reference set R representing the true underlying data manifold and an evaluation set E. If the sets R and E represent data from the same underlying manifold, then the geometric and topological properties extracted from manifolds described by R and E should be similar. DCA approximates these manifolds using a type of a neighbourhood graph called Delaunay graph $\mathcal{G}$ build on the union $R \cup E$. The alignment of R and E is then determined by analysing the connected components of $\mathcal{G}$ from which several global and local scores are derived.

DCA first evaluates each connected component $\mathcal{G}_i$ of $\mathcal{G}$ by analyzing the number of points from R and E contained in $\mathcal{G}_i$ as well as number of edges among these points. In particular, each component $\mathcal{G}_i$ is evaluated by two scores: *consistency* and *quality*. Intuitively, $\mathcal{G}_i$ has a high consistency if it is equality represented by points from R and E, and high quality if R and E points are geometrically well aligned. The latter holds true if the number of homogeneous edges among points in each of the sets is small compared to the number of heterogeneous edges connecting representations from R and E.

To formally define the scores, we follow Poklukar et al. (2022): for a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ we denote by $|\mathcal{G}|_\mathcal{V}$ the size of its vertex set and by $|\mathcal{G}|_\mathcal{E}$ the size of its edge set. Moreover, $\mathcal{G}^Q = (\mathcal{V}|_Q, \mathcal{E}|_{Q\times Q}) \subset \mathcal{G}$ denotes its restriction to a set $Q \subset \mathcal{V}$.

Definition A.1 *Consistency c and quality q of a connected component $\mathcal{G}_i \subset \mathcal{G}$ are defined as the ratios*

$$c(\mathcal{G}_i) = 1 - \frac{|\,|\mathcal{G}_i^R|_\mathcal{V} - |\mathcal{G}_i^E|_\mathcal{V}\,|}{|\mathcal{G}_i|_\mathcal{V}},$$
$$q(\mathcal{G}_i) = \begin{cases} 1 - \frac{(|\mathcal{G}_i^R|_\mathcal{E} + |\mathcal{G}_i^E|_\mathcal{E})}{|\mathcal{G}_i|_\mathcal{E}} & \text{if } |\mathcal{G}_i|_\mathcal{E} \geq 1 \\ 0 & \text{otherwise}, \end{cases}$$

respectively. Moreover, the scores computed on the entire Delaunay graph $\mathcal{G}$ are called network consistency $c(\mathcal{G})$ and network quality $q(\mathcal{G})$.

Besides the two global scores, *network consistency* and *network quality* defined above, two more global similarity scores are derived from the local ones by extracting the so-called *fundamental* components of high consistency and high quality. In this work, we define a component $\mathcal{G}_i$ to be fundamental if $c(\mathcal{G}_i) > 0$ and $q(\mathcal{G}_i) > 0$ and denote by $\mathcal{F}$ the union of all fundamental components of the Delaunay graph $\mathcal{G}$. By examining the proportion of points from E and R that are contained in $\mathcal{F}$, DCA derives two global scores *precision* and *recall* defined below.

Definition A.2 *Precision $\mathcal{P}$ and recall $\mathcal{R}$ associated to a Delaunay graph $\mathcal{G}$ built on $R \cup E$ are defined as*

$$\mathcal{P} = \frac{|\mathcal{F}^E|_\mathcal{V}}{|\mathcal{G}^E|_\mathcal{V}} \quad \text{and} \quad \mathcal{R} = \frac{|\mathcal{F}^R|_\mathcal{V}}{|\mathcal{G}^R|_\mathcal{V}},$$

respectively, where $\mathcal{F}^R, \mathcal{F}^E$ are the restrictions of $\mathcal{F}$ to the sets R and E.

We refer the reader to Poklukar et al. (2022; 2021) for further details.

B. Ablation Study on GMC

We perform a ablation study on the hyperparameters of GMC using the setup from Section 5.1 on the MHD dataset. In particular, we investigate:

1. the robustness of the GMC framework when varying the temperature parameter τ;
2. the performance of GMC with different dimensionalities of the intermediate representations $h \in \mathbb{R}^d$;
3. the performance of GMC with different dimensionalities of the shared latent representations $z \in \mathbb{R}^s$;
4. the performance of GMC with a modified loss $\mathcal{L}^*_{\text{GMC}}$ that only uses complete observations as negative pairs.

In all experiments we report both classification results and DCA scores.

Table 9. Performance of GMC with different temperature values τ (Equation (1)) in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$\tau = 0.05$	$\tau = 0.1$ (Default)	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.5$
Complete Observations	99.99 ± 0.01	100.00 ± 0.00	99.99 ± 0.01	$99.97 \pm 3e{-5}$	99.96 ± 0.01
Image Observations	99.78 ± 0.02	99.75 ± 0.03	99.84 ± 0.03	99.80 ± 0.04	99.89 ± 0.03
Sound Observations	93.55 ± 0.22	93.04 ± 0.45	91.98 ± 0.29	91.87 ± 0.58	95.01 ± 0.38
Trajectory Observations	99.94 ± 0.01	99.96 ± 0.02	99.97 ± 0.02	99.96 ± 0.01	99.80 ± 0.20
Label Observations	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table 10. DCA score obtained on GMC representations when trained with different temperature values τ (Equation (1)) 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	$\tau = 0.05$	$\tau = 0.1$ (Default)	$\tau = 0.2$	$\tau = 0.3$	$\tau = 0.5$
Complete ($z_{1:4}$)	Image (z_1)	0.96 ± 0.02	0.96 ± 0.02	0.93 ± 0.01	0.92 ± 0.00	0.89 ± 0.02
Complete ($z_{1:4}$)	Sound (z_2)	0.95 ± 0.02	0.87 ± 0.16	0.96 ± 0.02	0.99 ± 0.00	0.87 ± 0.04
Complete ($z_{1:4}$)	Trajectory (z_3)	0.96 ± 0.02	0.86 ± 0.05	0.90 ± 0.03	0.92 ± 0.00	0.64 ± 0.11
Complete ($z_{1:4}$)	Label (z_4)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.02

Table 11. Performance of GMC with different values of intermediate representation dimensionality $h \in \mathbb{R}^d$ in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete Observations	99.99 ± 0.01	100.00 ± 0.00	99.99 ± 0.01
Image Observations	99.75 ± 0.04	99.75 ± 0.03	99.72 ± 0.07
Sound Observations	93.31 ± 0.41	93.04 ± 0.45	93.34 ± 0.51
Trajectory Observations	99.96 ± 0.01	99.96 ± 0.02	99.96 ± 0.01
Label Observations	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table 12. DCA score obtained on GMC representations when varying the dimension of intermediate representations $h \in \mathbb{R}^d$ 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	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete ($z_{1:4}$)	Image (z_1)	0.91 ± 0.04	0.96 ± 0.02	0.92 ± 0.04
Complete ($z_{1:4}$)	Sound (z_2)	0.77 ± 0.17	0.87 ± 0.16	0.96 ± 0.04
Complete ($z_{1:4}$)	Trajectory (z_3)	0.86 ± 0.04	0.86 ± 0.05	0.86 ± 0.07
Complete ($z_{1:4}$)	Label (z_4)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Temperature parameter We study the performance of GMC when varying $\tau \in \{0.05, 0.1, 0.2, 0.3, 0.5\}$ (see Equation (1)). We present the classification results and DCA scores in Table 9 and Table 10, respectively. We observe that classification results are rather robust to different values of temperature, while increasing the temperature seems to have slightly negative effect on the geometry of the representations. For example, in Table 10, we observe that for $\tau = 0.5$ the trajectory representations z_3 are worse aligned with $z_{1:4}$.

Dimensionality of intermediate representations We vary the dimension of the intermediate representations space $d = \{32, 64, 128\}$ and present the resulting classification results and DCA scores in Table 11 and Table 12, respectively. The differences in classification results across different dimensions are covered by the margin of error, indicating the robustness of GMC to different sizes of the intermediate representations. We observe similar stability of the DCA scores in Table 10

Table 13. Performance of GMC with different values of latent representation dimensionality $z \in \mathbb{R}^s$ in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete Observations	99.99 ± 0.01	100.00 ± 0.00	99.99 ± 0.01
Image Observations	99.75 ± 0.04	99.75 ± 0.03	99.72 ± 0.07
Sound Observations	93.31 ± 0.41	93.04 ± 0.45	93.34 ± 0.51
Trajectory Observations	99.96 ± 0.01	99.96 ± 0.02	99.96 ± 0.01
Label Observations	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00

Table 14. DCA score obtained on GMC representations when varying the dimension of latent representations $z \in \mathbb{R}^d$ 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	$d = 32$	$d = 64$ (Default)	$d = 128$
Complete $(z_{1:4})$	Image (z_1)	0.93 ± 0.03	0.96 ± 0.02	0.91 ± 0.03
Complete $(z_{1:4})$	Sound (z_2)	0.89 ± 0.01	0.87 ± 0.16	0.86 ± 0.19
Complete $(z_{1:4})$	Trajectory (z_3)	0.81 ± 0.03	0.86 ± 0.05	0.88 ± 0.06
Complete $(z_{1:4})$	Label (z_4)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table 15. Performance of GMC with different loss functions in the MHD dataset, in a downstream classification task under complete and partial observations. Accuracy results averaged over 5 independent runs. Higher is better.

Observations	$\mathcal{L}_{\text{GMC}}$ (Default)	$\mathcal{L}^*_{\text{GMC}}$
Complete Observations	100.00 ± 0.00	99.97 ± 0.02
Image Observations	99.75 ± 0.03	99.87 ± 0.01
Sound Observations	93.04 ± 0.45	92.79 ± 0.24
Trajectory Observations	99.96 ± 0.02	99.98 ± 0.01
Label Observations	100.00 ± 0.00	100.00 ± 0.00

Table 16. DCA score obtained on GMC representations when trained different loss functions 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	$\mathcal{L}_{\text{GMC}}$ (Default)	$\mathcal{L}^*_{\text{GMC}}$
Complete $(z_{1:4})$	Image (z_1)	0.96 ± 0.02	0.80 ± 0.02
Complete $(z_{1:4})$	Sound (z_2)	0.87 ± 0.16	0.27 ± 0.14
Complete $(z_{1:4})$	Trajectory (z_3)	0.86 ± 0.05	0.86 ± 0.03
Complete $(z_{1:4})$	Label (z_4)	1.00 ± 0.00	0.24 ± 0.10

with minor variations in the geometric alignment for the sound modality z_2 which benefits from the larger intermediate representation space.

Dimensionality of latent representations We repeat a similar evaluation for the dimension of the latent space $s = \{32, 64, 128\}$ and present the classification and DCA scores in Table 13 and Table 14, respectively. We observe that GMC is robust to changes in s both in terms of performance and geometric alignment.

Loss function We consider an ablated version of the loss function, $\mathcal{L}^*_{\text{GMC}}$, that considers only complete-observations as negative pair, i.e. $\Omega^*(i) = s_{1:M,1:M}(i, j)$ for $j = 1, \ldots, B$ where B is the size of the mini-batch. We present the classification results and DCA scores in Table 15 and Table 16, respectively. The results in Table 15 highlight the importance of the contrasting the complete representations to learn a robust representation suitable for downstream tasks as we observe minimal variation in classification accuracy when considering different loss. However, we observe worse geometric

Table 17. Performance of different multimodal representation methods in the CMU-MOSI 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 (↓)	1.033 ± 0.037	$\mathbf{1.010 \pm 0.070}$
Cor (↑)	0.642 ± 0.008	$\mathbf{0.649 \pm 0.019}$
F1 (↑)	0.770 ± 0.017	$\mathbf{0.776 \pm 0.023}$
Acc (%, ↑)	77.07 ± 01.67	$\mathbf{77.59 \pm 02.20}$

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

Metric	Baseline	GMC (Ours)
MAE (↓)	1.244 ± 0.100	$\mathbf{1.119 \pm 0.033}$
Cor (↑)	0.431 ± 0.208	$\mathbf{0.573 \pm 0.016}$
F1 (↑)	0.698 ± 0.053	$\mathbf{0.727 \pm 0.013}$
Acc (%, ↑)	66.28 ± 07.74	$\mathbf{72.32 \pm 0.013}$

(b) Text Observations (x_1)

Metric	Baseline	GMC (Ours)
MAE (↓)	$\mathbf{1.431 \pm 0.025}$	1.434 ± 0.017
Cor (↑)	0.056 ± 0.071	$\mathbf{0.211 \pm 0.010}$
F1 (↑)	$\mathbf{0.588 \pm 0.076}$	0.570 ± 0.006
Acc (%, ↑)	47.20 ± 05.67	$\mathbf{55.91 \pm 01.11}$

(c) Audio Observations (x_2)

Metric	Baseline	GMC (Ours)
MAE (↓)	$\mathbf{1.406 \pm 0.041}$	1.452 ± 0.035
Cor (↑)	0.021 ± 0.028	$\mathbf{0.176 \pm 0.028}$
F1 (↑)	$\mathbf{0.659 \pm 0.049}$	0.550 ± 0.015
Acc (%, ↑)	53.87 ± 05.77	$\mathbf{54.30 \pm 01.96}$

(d) Video Observations (x_3)

Table 18. DCA score of the models in the CMU-MOSI dataset, evaluating the geometric alignment of complete representations $z_{1:4}$ and modality-specific ones $\{z_1, z_2, z_3\}$ used as R and E inputs in DCA, respectively. The score is averaged over 5 independent runs. Higher is better.

R	E	Baseline	**GMC (Ours)**
Complete $(z_{1:3})$	Text (z_1)	0.54 ± 0.07	$\mathbf{0.93 \pm 0.02}$
Complete $(z_{1:3})$	Audio (z_2)	0.14 ± 0.06	$\mathbf{0.75 \pm 0.05}$
Complete $(z_{1:3})$	Vision (z_3)	0.36 ± 0.09	$\mathbf{0.85 \pm 0.04}$

alignment when using $\mathcal{L}^*_{\text{GMC}}$ loss during training of GMC. This suggests that contrasting among individual modalities is beneficial for geometrical alignment of the representations.

C. Experiment 2: Supervised Learning with the CMU-MOSI dataset

In this section, we repeat the experimental evaluation of Section 5.2 with the CMU-MOSI dataset. We employ the same baseline and GMC architectures as in the CMU-MOSEI evaluation and consider the same evaluation setup.

Results The results obtained on CMU-MOSI are reported in Table 17. We observe that GMC improves the robustness of the model to the missing modalities as seen from Tables 17b, 17c and 17d where we use only individual modalities as inputs. However, the increase in performance is not as signification as in the case of the CMU-MOSEI dataset for audio (x_2) and video (x_3) modalities where the baseline outperforms GMC on MAE and F1 scores. We hypothesise that this behaviour is due to the intrinsic difficulty of forming good contrastive pairs in small-sizes datasets (Cao & Wu, 2021): the CMU-MOSI dataset has only 1513 training samples which hinders the learning of a quality latent representations. However, we observe that GMC still significantly improves the geometric alignment (Table 18) of the modality-specific representations z_m (comprising the set E) and complete representations $z_{1:3}$ (comprising the set R) compared to the baseline, even in this regime of small data.

D. Model Architecture

We report the model architectures for GMC employed in our work: in Figure 4 we present the model employed for the unsupervised experiment of Section 5.1; in Figure 5 we present the model employed for the supervised experiment of Section 5.2; in Figure 6 we present the model employed in the RL experiment of Section 5.3.

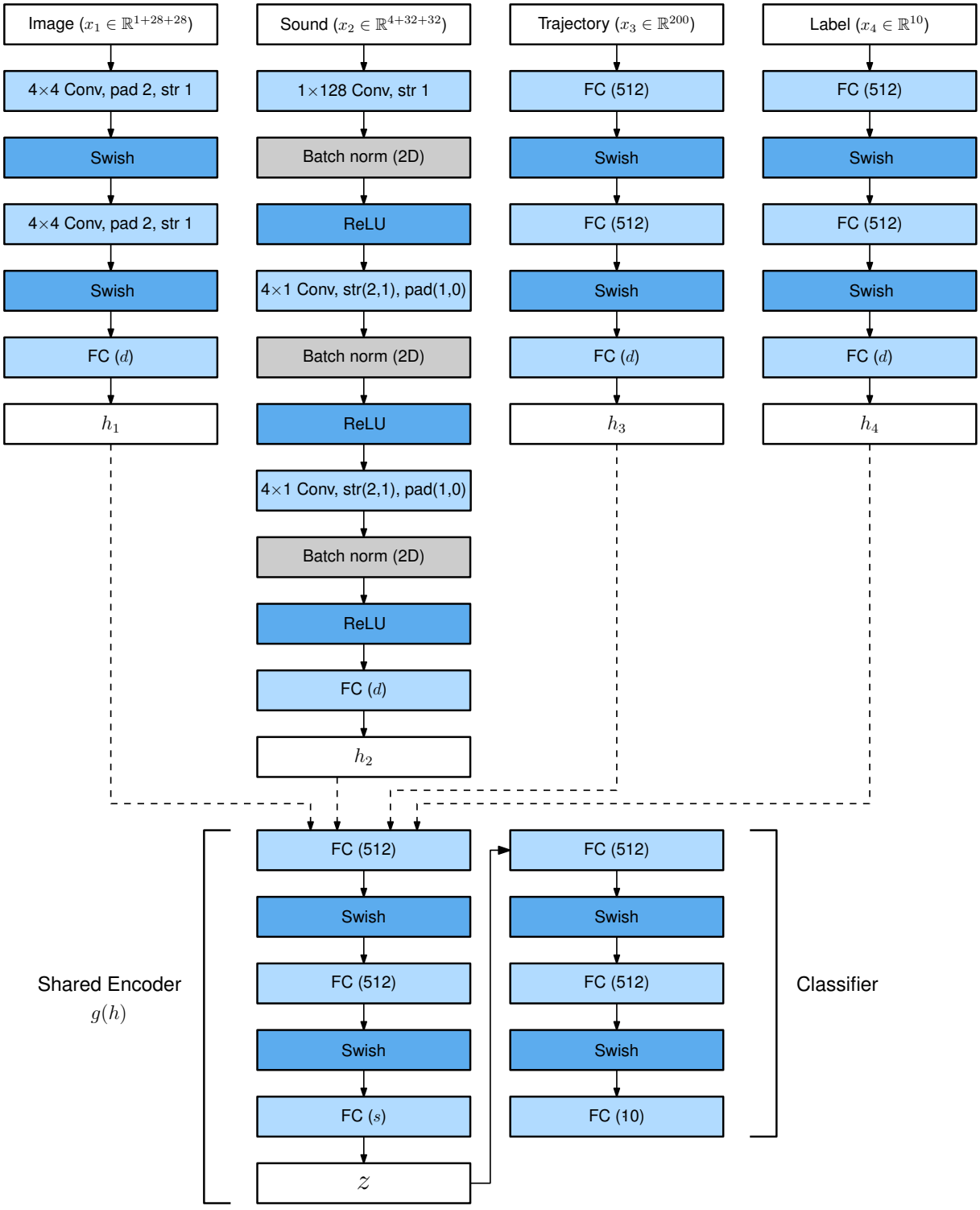


Figure 4. GMC model for the unsupervised experiment of Section 5.1. Dashed lines represent potential connections between the intermediate representations $\{h_1, \ldots, h_4\}$ and the shared head $g(h)$. For the joint modality base encoder (not depicted) we employ an additional network with an identical architecture to the modality-specific ones, employing a late-fusion mechanism of all modalities before the projection (FC) to the intermediate representation h.

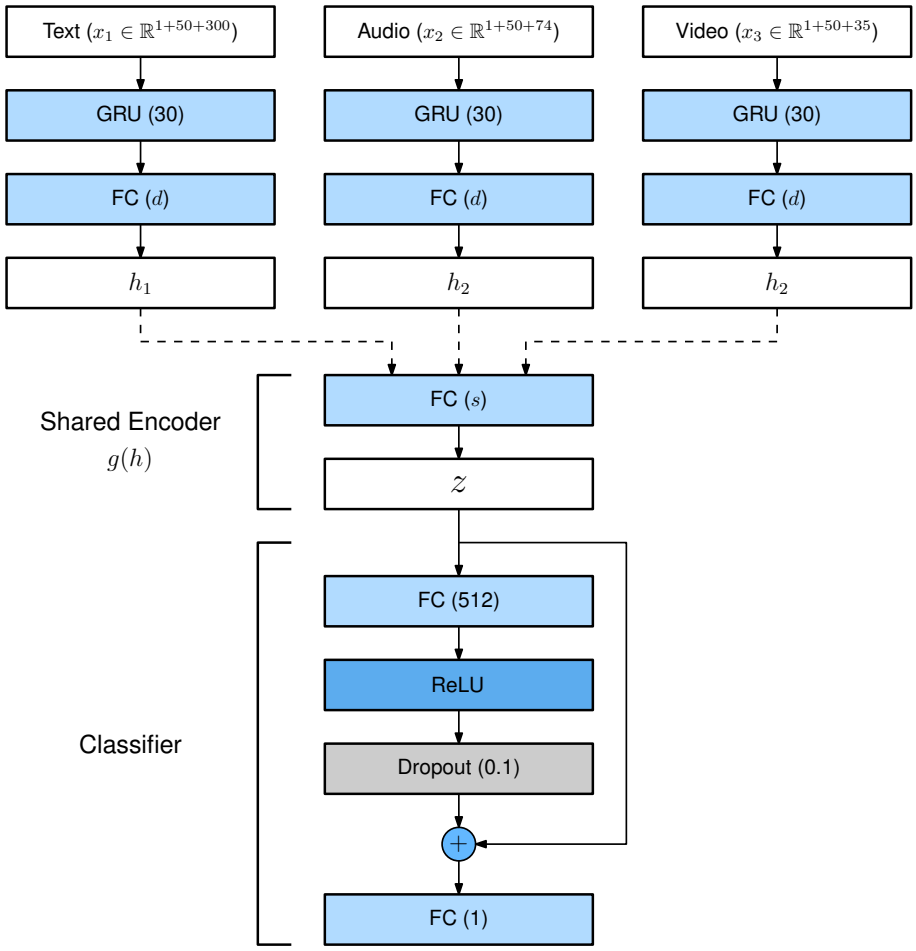


Figure 5. GMC model for the supervised experiment of Section 5.2. Dashed lines represent potential connections between the intermediate representations $\{h_1, \ldots, h_3\}$ and the shared head $g(h)$. For the joint modality base encoder (not depicted) we employ the baseline multimodal transformer model, whose architecture we refer to Tsai et al. (2019a).

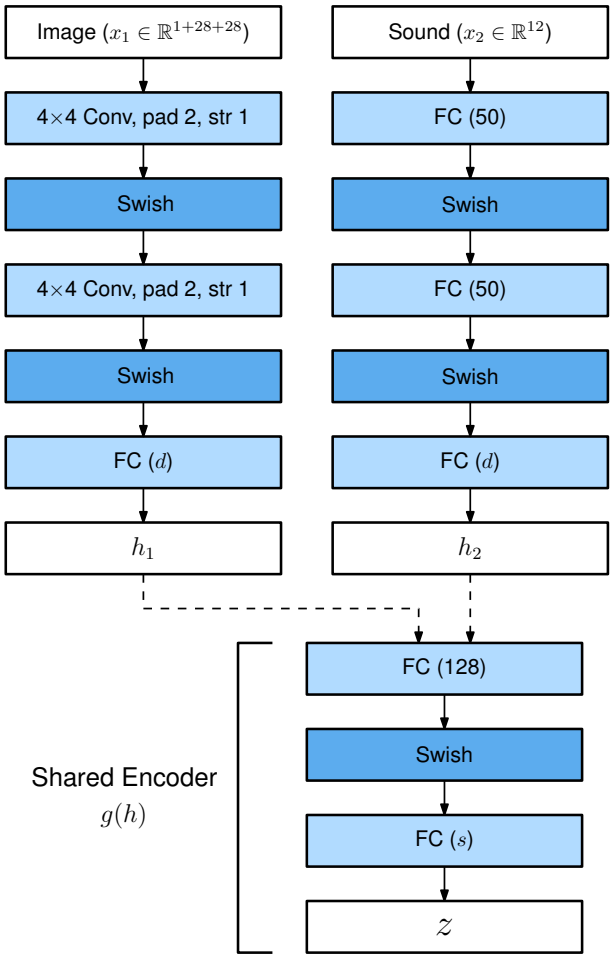
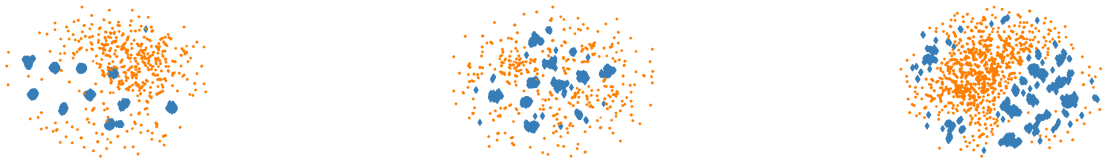
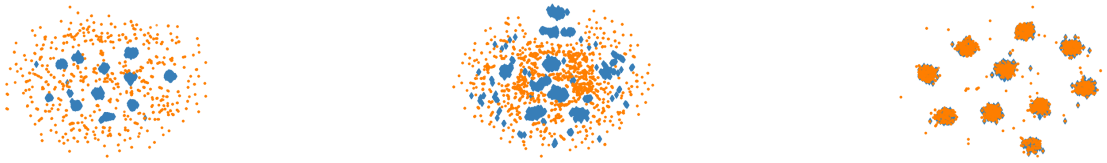


Figure 6. GMC model for the RL experiment of Section 5.3. Dashed lines represent potential connections between the intermediate representations $\{h_1, h_2\}$ and the shared head $g(h)$. For the joint modality base encoder (not depicted) we employ an additional network with an identical architecture to the modality-specific ones, employing a late-fusion mechanism of all modalities before the projection (FC) to the intermediate representation h. For the policy network, we refer to Silva et al. (2020).



(a) MVAE (b) MMVAE (c) Nexus



(d) MUSE (e) MFM (f) **GMC (Ours)**

Figure 7. UMAP visualization of complete representations $z_{1:4}$ (blue) and sound representations z_2 (orange) obtained from several state-of-the-art multimodal representation learning models on the MHD dataset considered in Section 5.1. Best viewed in color.

E. Training Hyperparameters

In Table 19 we present the hyperparameters employed in this work. For training the controller in the RL task, we employ the same training hyperparameters as in Silva et al. (2020).

F. Additional Visualizations of the Alignment of Complete and Modality-Specific Representations

We present additional visualizations of encodings of complete and modality-specific representations in the MHD dataset for multiple multimodal representation models. In Figures 7, 8 and 9, we show visualizations of sound representations z_2, trajectory z_3 and label z_4 (in orange), respectively, and complete representations $z_{1:4}$ (in blue). Note that points detected as outliers by DCA are *not* included in the visualization. For example, we observe that certain labels representations for baseline models are marked as outliers in Figure 9.

Table 19. Training hyperparameters of GMC.

(a) Unsupervised (Section 5.1)

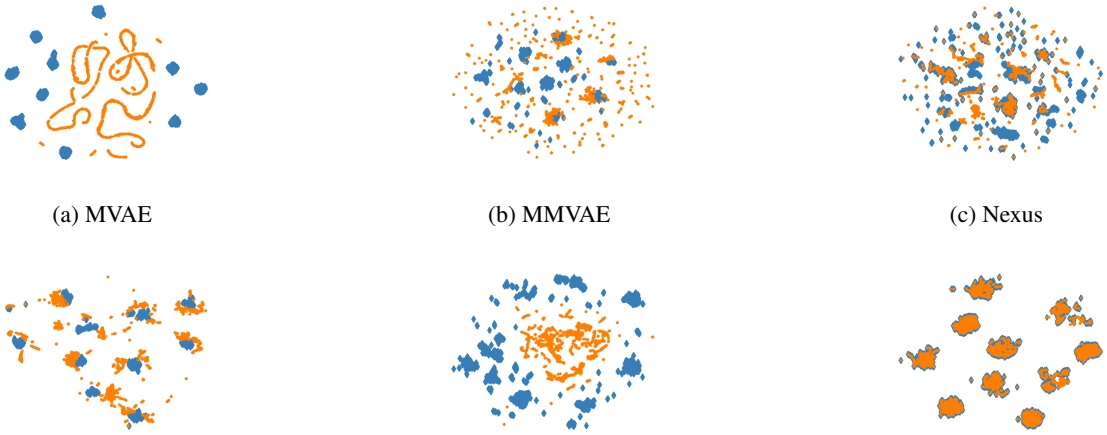
Parameter	Value
Intermediate size d	64
Latent size s	64
Model training epochs	100
Classifier training epochs	50
Learning rate	1e−3
Batch size B	64
Temperature τ	0.1

(b) Supervised (Section 5.2)

Parameter	Value
Intermediate size d	60
Latent size s	60
Model training epochs	40
Learning rate	1e−3 (Decay)
Batch size B	40
Temperature τ	0.3

(c) RL (Section 5.3)

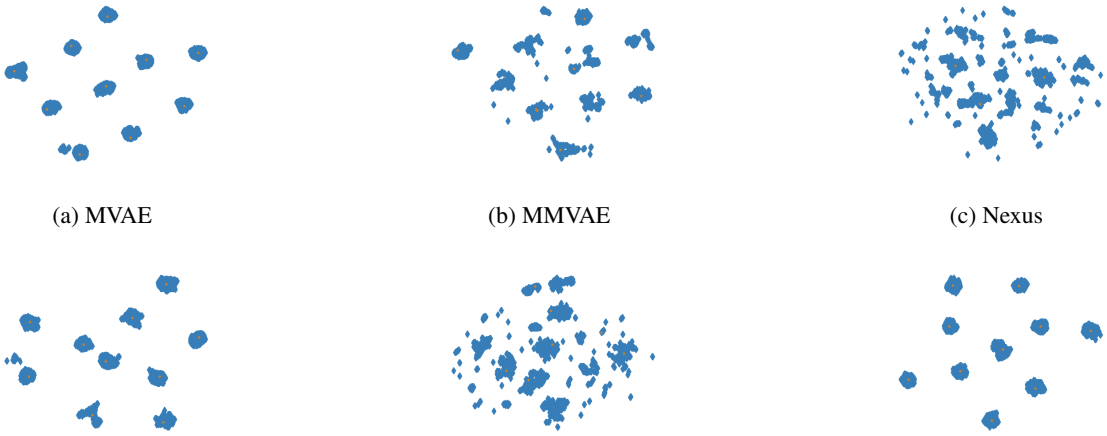
Parameter	Value
Intermediate size d	64
Latent size s	10
Model training epochs	500
Learning rate	1e−3 (Decay)
Batch size B	128
Temperature τ	0.3



(a) MVAE (b) MMVAE (c) Nexus

(d) MUSE (e) MFM (f) **GMC (Ours)**

Figure 8. UMAP visualization of complete representations $z_{1:4}$ (blue) and trajectory representations z_3 (orange) obtained from several state-of-the-art multimodal representation learning models on the MHD dataset considered in Section 5.1. Best viewed in color.



(a) MVAE (b) MMVAE (c) Nexus

(d) MUSE (e) MFM (f) **GMC (Ours)**

Figure 9. UMAP visualization of complete representations $z_{1:4}$ (blue) and label representations z_4 (orange) obtained from several state-of-the-art multimodal representation learning models on the MHD dataset considered in Section 5.1. Best viewed in color.