Technical Appendix

Datasets

- **Clustering Datasets.** We conduct the clustering experiments on four widely used multi-modal dataset to evaluate our method. The multi-modal samples in each dataset are well-aligned and have the corresponding labels, which are only used in the validation stage. The links of these four clustering dataset from (Hu, Nie, and Li 2019; Guo et al. 2014; Gong, Nie, and Xu 2022) are given as follows:
  - **Caltech7** https://data.caltech.edu/records/20086
  - **ORL** https://cam-orl.co.uk/facedatabase.html
  - **Movies** http://membres-lig.imag.fr/grimal/data.html
  - **Prokaryotic** https://github.com/mbrbic/Multi-view-LRSSC/blob/master/datasets
- **Classification & Regression Datasets.** The dataset for the classification and regression tasks are chosen from Multibench (Liang et al. 2021), and are pre-processed and filtered in the same way as Multibench does. Detailed information of all dataset can be found in section C of Multibench (Liang et al. 2021)'s Appendix.
  - **CMU-MOSI** (Zadeh et al. 2016) CMU-MOSI is used in affective computing areas, which contains language, video, and audio time-series data, to predict sentiment. According to Multibench (Liang et al. 2021), CMU-MOSI was originally downloaded from https://github.com/A2Zadeh/CMU-MultimodalSDK.
  - **MUJOCO** (Lee et al. 2020) The MUJOCO dataset we use in this paper is short for MUJOCO PUSH, which is a large-scale dataset in robotics to predict the pose of the object being pushed by the robot end-effector (Liang et al. 2021). It has records on the manipulation of simulated and real robotic arms equipped with visual (RGB and depth), force, and proprioception sensors. It can be downloaded from https://github.com/brentyi/multimodalfilter/.
  - **ENRICO** (Leiva, Hota, and Oulasvirta 2020) ENRICO (Enhanced Rico) dataset is a benchmarks for data-driven models of design in scaffolding the creation of mobile apps (Liang et al. 2021), which consists of two modalities for app classification: (1) the app screenshot image and (2) the set of unordered view hierarchy, which describes the spatial and structural layout of UI elements. It can be downloaded from https://github.com/luileito/enrico.
  - **AV-MNIST** (Vielzeuf et al. 2018) AV-MNIST pairs the audio of a human reading digits from the FSDD dataset in https://github.com/Jakobovski/free-spoken-digit-dataset with written digits in the MNIST dataset (LeCun et al. 1998), and the task is to predict the class of the digit from 0 to 9. According to Multibench (Liang et al. 2021), the preprocessing code is provided in https://github.com/slyviacassell/_MFAS/blob/master/datasets/avmnist_gen.py.

Metrics

- **Clustering Metric.** Purity is a traditional metric to measure the quality of clustering task and is widely used (Sripada and Rao 2011; Marutho et al. 2018). Purity value is computed by equation (1). The purity value is no more than 1, and the higher the purity is, the better of the clustering performance.

$$\text{Purity} = \frac{1}{N} \sum_{k=1}^{K} \max_{j} |C_k \cap L_j| \tag{1}$$

  where N is the total number of samples, K is the number of clusters, C_k is the k-th cluster, L_j is the set of samples with true label j, $|C_k \cap L_j|$ is the number of samples in the k-th cluster that belong to the true label j.
- **Classification and Regression Metric.** Accuracy (Aronoff et al. 1982; Baldi et al. 2000) is a traditional metric to measure the quality of classification task, while MSE (Willmott and Matsuura 2005) is a traditional metric to measure the quality of regression task. For regression, MAE is better than MSE for evaluating the performance as stated in (Willmott and Matsuura 2005). The less MSE or MAE is, the better of the regression performance.

Implementation Details

The random seed is setup at the beginning of the experiment, which can be found in the code files. For classification and regression experiments, the initial seed is 42. For clustering experiments, the initial seed is 123. We run our experiments on two sets of computing infrastructures. One has CPU Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz, GPU model NVIDIA RTX A6000 with OS 20.04.1-Ubuntu x86_64 GNU, and the amount of memory is 251Gi. The other one has CPU Intel(R) Xeon(R) Gold 5218 CPU @ 2.30GHz, GPU NVIDIA GeForce RTX 3090, with OS 83$\sim$20.04.1-Ubuntu x86_64 GNU/Linux, and the amount of memory is 251Gi. The relevant software libraries and frameworks are given in the code repositories by the yml files.

- **Clustering.** The details of the parameters and models can be found in the config directory of the repository name mwae. We conduct the clustering training and validation on the full dataset as the baselines did. The optimizer is Adam, the early stop is adopted with patience 5. α is the hyper-parameter of FGW, which varies from 0 to 1, and we tune it in [0,1] by uniform with interval 0.095333. The weight of OTM is tuned in [1e-2, 1e2] by loguniform. The weight of reconstruction loss is tuned in [1e-2, 1e2] by loguniform. The learning rate is tuned in [1e-5, 5e-2] by loguniform.
- **Classification.** The hyper-parameters of the classification and regression tasks are listed in Table 2. We only tune the OTM weight in [1e-2,1e2] by loguniform, and other parameters are consistent with Multibench. For CMU-MOSI, the max number of epochs is 100, the batch size is 32, the early stop is not applied except for the LRTF with patience 7, the optimizer is AdamW, the weight decay is 0.01, and the learning rate is 0.001. For

Dataset	Aligned	α	Construction Weight	OTM Weight
Caltech7	False	0.8579	98.6208	0.4575
	True	0.0953	1.6462	0.0138
Movies	False	0.0953	47.2208	0.1701
	True	0.5719	33.6095	16.1002
ORL	False	0.7626	99.0937	0.1391
	True	0.4766	15.2616	15.6488
Prokaryotic	False	0.1906	1.6734	0.7819
	True	0.7626	0.0232	5.6824

Table 1: Hyper-parameter for clustering task on various dataset.

ENRICO, the max number of epochs is 50, the batch size is 32, the early stop is not applied, the optimizer is Adam, the learning rate is 0.0001, and there is no weight decay. For AV-MNIST, the max number of epochs is 30, the batch size is 32, the early stop is not applied, the optimizer is SGD, the weight decay is 0.0001, and the learning rate is 0.1.

- **Regression.** For CMU-MOSI, the max number of epochs is 100, the batch size is 32, the early stop is not applied, the optimizer is AdamW, the weight decay is 0.01 and the learning rate is 0.001. For MUJOCO, the max number of epochs is 20, the batch size is 32, the early stop is not applied, the opimizer is AdamW, the learning rate is 0.00001, and there is no weight decay. We only tune the OTM weight in [1e-2,1e2] by loguniform, and other parameters are consistent with Multibench.
- **Robustness to missing modalities.** For Prokaryotic, we use the same settings of the multi-modal training to train each modality one by one. For CMU-MOSI, we loaded the model learnt with or without OTM in the previous classification and regression tasks. Then we select the dimension of the single modality's head to match the dimension of the model's output to perform classification or regression tasks. Finally, we train the head by means of the single modality training data, and valid the models one modality by one modality on the single modality validation data by 5-fold cross-validation. We implement the head by MLP, and the dimensions of the head are listed in Table 3.
- **Robustness to data noise.** For CMU-MOSI and MUJOCO dataset, we add noise to all modalities with different noise levels around 0.2 for robustness comparison. The hyper-parameters are listed in 4. The validation task for classification task on CMU-MOSI is classification, while the validation task for regression task on CMU-MOSI is posneg-classification, which is consistent with Multibench. The validation task for regression task on MUJOCO is regression. For this experiment, we tune the loss level in [0.1, 0.3] with interval 0.1.

Dataset	Task	Fusion	OTM weight
CMU-MOSI	Classification	Late fusion	100.0
		LRTF	0.01
		MFM	1.0
	Regression	Late fusion	100.0
		TF	10.0
ENRICO	Classification	MI matrix	0.1
		Tensor matrix	100.0
AV-MNIST	Classification	Late fusion	0.01
MUJOCO	regression	TF	0.1

Table 2: Hyper-parameter for classification and regression tasks on various dataset.

Task	Fusion	Modality	Head
Classification	Late fusion	1	[70, 870, 2]
		2	[200, 870, 2]
		3	[600, 870, 2]
	LRTF	1	[128, 512, 2]
		2	[128, 512, 2]
		3	[128, 512, 2]
Regressioin	Late fusion	1	[70, 870, 1]
		2	[200, 870, 1]
		3	[600, 870, 1]
	Tensor Fusion	1	[4,4,1]
		2	[19,19,1]
		3	[79,79,1]

Table 3: Head parameters of CMU-MOSI's single modality.

References

Aronoff, S.; et al. 1982. Classification accuracy: a user approach. *Photogrammetric Engineering and Remote Sensing*, 48(8): 1299–1307.

Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; and Nielsen, H. 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5): 412–424.

Gong, F.; Nie, Y.; and Xu, H. 2022. Gromov-Wasserstein multi-modal alignment and clustering. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, 603–613.

Guo, D.; Zhang, J.; Liu, X.; Cui, Y.; and Zhao, C. 2014. Multiple kernel learning based multi-view spectral clustering. In *2014 22nd International conference on pattern recognition*, 3774–3779. IEEE.

Hu, D.; Nie, F.; and Li, X. 2019. Deep multimodal clustering for unsupervised audiovisual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9248–9257.

LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11): 2278–2324.

Lee, M. A.; Yi, B.; Martín-Martín, R.; Savarese, S.; and Bohg, J. 2020. Multimodal sensor fusion with differentiable filters. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 10444–10451. IEEE.

Dataset	Train Task	Fusion	Noise	β	Epoch
CMU-MOSI	Classification	Late fusion	0.2	100.0	300
		LRTF	0.3	0.01	100
		MFM	0.2	0.1	100
	Regressioin	Late fusion	0.2	100.0	100
		Tensor Fusion	0.2	10.0	300
MUJOCO	Regressioin	Tensor Fusion	0.3	0.1	20

Table 4: Hyper-parameters of noisy test on CMU-MOSI and MUJOCO dataset.

Leiva, L. A.; Hota, A.; and Oulasvirta, A. 2020. Enrico: A high-quality dataset for topic modeling of mobile UI designs. *Proc. MobileHCI extended abstracts*.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*.

Marutho, D.; Handaka, S. H.; Wijaya, E.; et al. 2018. The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 international seminar on application for technology of information and communication*, 533–538. IEEE.

Sripada, S. C.; and Rao, M. S. 2011. Comparison of purity and entropy of k-means clustering and fuzzy c means clustering. *Indian journal of computer science and engineering*, 2(3): 343–346.

Vielzeuf, V.; Lechervy, A.; Pateux, S.; and Jurie, F. 2018. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Willmott, C. J.; and Matsuura, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research*, 30(1): 79–82.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.