Learning Multimodal Representations with Factorized Deep Generative Models

Yao-Hung Hubert Tsai[*1], **Paul Pu Liang**[*1],
Amir Zadeh[2], **Louis-Philippe Morency**[2], **Ruslan Salakhutdinov**[1]
{[1]Machine Learning Department, [2]Language Technologies Institute}, Carnegie Mellon University
{yaohungt,pliang,abagherz,morency,rsalakhu}@cs.cmu.edu

Abstract

Learning multimodal representations is a fundamentally complex research problem due to the presence of multiple heterogeneous sources of information. Although the presence of multiple modalities provides additional valuable information, there are two key challenges to address when learning from multimodal data: 1) models must learn the complex intra-modal and cross-modal interactions for prediction and 2) models must be robust to unexpected missing or noisy modalities during testing. In this paper, we propose to optimize for a joint generative-discriminative objective across multimodal data and labels. We introduce a model that factorizes representations into two sets of independent factors: *multimodal discriminative* and *modality-specific generative* factors. Multimodal discriminative factors are shared across all modalities and contain joint multimodal features required for discriminative tasks such as sentiment prediction. Modality-specific generative factors are unique for each modality and contain the information required for generating data. Experimental results show that our model is able to learn meaningful multimodal representations that achieve state-of-the-art or competitive performance on six multimodal datasets. Our model demonstrates flexible generative capabilities by conditioning on independent factors and can reconstruct missing modalities without significantly impacting performance. Lastly, we interpret our factorized representations to understand the interactions that influence multimodal learning.

1 Introduction

Multimodal machine learning involves learning from data across multiple modalities [1]. It is a challenging yet crucial research area with real-world applications in robotics [33], dialogue systems [46], intelligent tutoring systems [44], and healthcare diagnosis [15]. At the heart of many multimodal modeling tasks lies the challenge of learning rich representations from multiple modalities. For example, analyzing multimedia content requires learning multimodal representations across the language, visual, and acoustic modalities [9]. Although the presence of multiple modalities provides additional valuable information, there are two key challenges to address when learning from multimodal data: 1) models must learn the complex intra-modal and cross-modal interactions for prediction [67], and 2) trained models must be robust to unexpected missing or noisy modalities during testing [39].

In this paper, we propose to optimize for a joint generative-discriminative objective across multimodal data and labels. The discriminative objective ensures that the representations learned are rich in intra-modal and cross-modal features useful towards predicting the label, while the generative objective allows the model to infer missing modalities at test time and deal with the presence of noisy modalities. To this end, we introduce the Multimodal Factorization Model (MFM in Figure 1) that factorizes multimodal representations into *multimodal discriminative* factors and *modality-specific*

[*]equal contributions

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

generative factors. Multimodal discriminative factors are shared across all modalities and contain joint multimodal features required for discriminative tasks. Modality-specific generative factors are unique for each modality and contain the information required for generating each modality. We believe that factorizing multimodal representations into different explanatory factors can help each factor focus on learning from a subset of the joint information across multimodal data and labels. This method is in contrast to jointly learning a single factor that summarizes all generative and discriminative information [57]. To sum up, MFM defines a joint distribution over multimodal data, and by the conditional independence assumptions in the assumed graphical model, both generative and discriminative aspects are taken into account. Our model design further provides interpretability of the factorized representations.

Through an extensive set of experiments, we show that MFM learns improved multimodal representations with these characteristics: 1) The multimodal discriminative factors achieve state-of-the-art or competitive performance on six multimodal time series datasets. We also demonstrate that MFM can generalize by integrating it with other existing multimodal discriminative models. 2) MFM allows flexible generation concerning multimodal discriminative factors (labels) and modality-specific generative factors (styles). We further show that we can perform reconstruction of missing modalities from observed modalities without significantly impacting discriminative performance. Finally, we interpret our learned representations using information-based and gradient-based methods, allowing us to understand the contributions of individual factors towards multimodal prediction and generation.

2 Multimodal Factorization Model

Multimodal Factorization Model (MFM) is a latent variable model (Figure 1(a)) with conditional independence assumptions over multimodal discriminative factors and modality-specific generative factors. We propose a factorization over the joint distribution of multimodal data (Section 2.1). Since exact posterior inference on this factorized distribution can be intractable, we propose an approximate inference algorithm based on minimizing a joint-distribution Wasserstein distance over multimodal data (Section 2.2). Finally, we derive the MFM objective by approximating the joint-distribution Wasserstein distance via a generalized mean-field assumption.

Notation: Define $\mathbf{X}_{1:M}$ as multimodal data from M modalities and $\mathbf{Y}$ as the labels, with joint distribution $P_{\mathbf{X}_{1:M},\mathbf{Y}} = P(\mathbf{X}_{1:M}, \mathbf{Y})$. $\hat{\mathbf{X}}_{1:M}$ denotes the generated data and $\hat{\mathbf{Y}}$ the generated labels, with joint distribution $P_{\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}}}$.

2.1 Factorized Multimodal Representations

To factorize multimodal representations into multimodal discriminative factors and modality-specific generative factors, MFM assumes a Bayesian network structure as shown in Figure 1(a). In this graphical model, factors $\mathbf{F_y}$ and $\mathbf{F_{a\{1:M\}}}$ are generated from mutually independent latent variables $\mathbf{Z} = [\mathbf{Z_y}, \mathbf{Z_{a\{1:M\}}}]$ with prior $P_{\mathbf{Z}}$. In particular, $\mathbf{Z_y}$ generates the multimodal discriminative factor $\mathbf{F_y}$ and $\mathbf{Z_{a\{1:M\}}}$ generate modality-specific generative factors $\mathbf{F_{a\{1:M\}}}$. By construction, $\mathbf{F_y}$ contributes to the generation of $\hat{\mathbf{Y}}$ while $\{\mathbf{F_y}, \mathbf{F_{a}}_i\}$ both contribute to the generation of $\hat{\mathbf{X}}_i$. As a result, the joint distribution $P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}})$ can be factorized as follows:

$$
\begin{aligned}
P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}}) &= \int_{\mathbf{F},\mathbf{Z}} P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}}|\mathbf{F}) P(\mathbf{F}|\mathbf{Z}) P(\mathbf{Z}) d\mathbf{F} d\mathbf{Z} \\
&= \int_{\substack{\mathbf{F_y},\mathbf{F_{a\{1:M\}}} \\ \mathbf{Z_y},\mathbf{Z_{a\{1:M\}}}}} \Big(P(\hat{\mathbf{Y}}|\mathbf{F_y}) \prod_{i=1}^{M} P(\hat{\mathbf{X}}_i|\mathbf{F_{a}}_i, \mathbf{F_y})\Big)\Big(P(\mathbf{F_y}|\mathbf{Z_y}) \prod_{i=1}^{M} P(\mathbf{F_{a}}_i|\mathbf{Z_{a}}_i)\Big)\Big(P(\mathbf{Z_y}) \prod_{i=1}^{M} P(\mathbf{Z_{a}}_i)\Big) d\mathbf{F} d\mathbf{Z},
\end{aligned} \tag{1}
$$

with $d\mathbf{F} = d\mathbf{F_y} \prod_{i=1}^{M} d\mathbf{F_{a}}_i$ and $d\mathbf{Z} = d\mathbf{Z_y} \prod_{i=1}^{M} d\mathbf{Z_{a}}_i$. Exact posterior inference in Equation 1 may be analytically intractable due to the integration over $\mathbf{Z}$. We therefore use an approximate inference distribution $Q(\mathbf{Z}|\mathbf{X}_{1:M}, \mathbf{Y})$. MFM can be viewed as an autoencoding structure that consists of encoder (inference) and decoder (generative) modules (Figure 1(c)). The encoder module for $Q(\cdot|\cdot)$ allows us to easily sample $\mathbf{Z}$ from an approximate posterior. The decoder modules are parametrized according to the factorization of $P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}}|\mathbf{Z})$ as given by Equation 1 and Figure 1(a).

2.2 Minimizing Joint-Distribution Wasserstein Distance over Multimodal Data

Two common choices for approximate inference in autoencoding structures are Variational Autoencoders (VAEs) [26] and Wasserstein Autoencoders (WAEs) [71, 60]. The former optimizes the

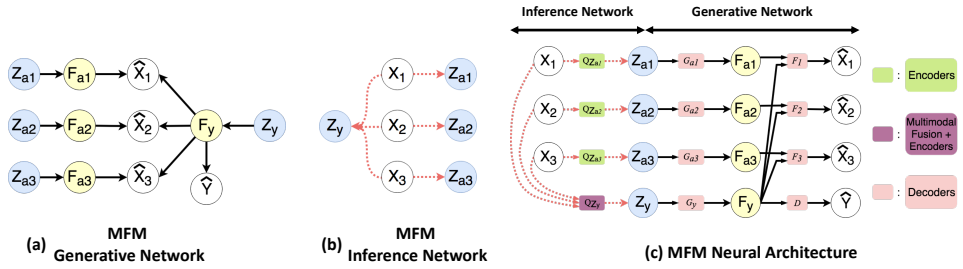


Figure 1: Illustration of the proposed Multimodal Factorization Model (MFM) with three modalities. MFM factorizes multimodal representations into *multimodal discriminative* factors $\mathbf{F_y}$ and *modality-specific generative* factors $\mathbf{F_{a\{1:M\}}}$. (a) MFM Generative Network with latent variables $\{\mathbf{Z_y}, \mathbf{Z_{a\{1:M\}}}\}$, factors $\{\mathbf{F_y}, \mathbf{F_{a\{1:M\}}}\}$, generated multimodal data $\hat{\mathbf{X}}_{1:3}$ and labels $\hat{\mathbf{Y}}$. (b) MFM Inference Network. (c) MFM Neural Architecture.

evidence lower bound objective (ELBO), and the latter derives an approximation for the primal form of the Wasserstein distance. Wasserstein Autoencoders (WAEs) [71, 60] have been a popular choice for approximate inference in autoencoding structures since they perform better latent factor disentanglement [71, 51] and better sample generation quality [7, 21, 26]. However, WAEs are designed for unimodal data and do not consider factorized distributions over latent variables that generate multimodal data. We propose a variant of WAEs to handle factorized joint distributions over multimodal data. We begin by using neural networks in the encoder and decoder [26] (Figure 1 (c)). For the encoder $Q(\mathbf{Z}|\mathbf{X}_{1:M}, \mathbf{Y})$, we learn a deterministic mapping $Q_{enc} : \mathbf{X}_{1:M}, \mathbf{Y} \rightarrow \mathbf{Z}$ [51, 60]. For the decoder, we define the generation process from latent variables as $G_y : \mathbf{Z_y} \rightarrow \mathbf{F_y}$, $G_{a\{1:M\}} : \mathbf{Z_{a\{1:M\}}} \rightarrow \mathbf{F_{a\{1:M\}}}$, $D : \mathbf{F_y} \rightarrow \hat{\mathbf{Y}}$, and $F_{1:M} : \mathbf{F_y}, \mathbf{F_{a\{1:M\}}} \rightarrow \hat{\mathbf{X}}_{1:M}$, where $G_y, G_{a\{1:M\}}, D$ and $F_{1:M}$ are deterministic functions parametrized by neural networks.

Let $W_c(P_{\mathbf{X}_{1:M},\mathbf{Y}}, P_{\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}}})$ denote the joint-distribution Wasserstein distance over multimodal data under cost function $c_{X\,i}$ and c_Y. We choose the squared cost $c(a, b) = \|a - b\|_2^2$, allowing us to minimize the 2-Wasserstein distance. The cost function can be defined not only on static data but also on time series data such as text, audio and videos. For example, given time series data $\mathbf{X} = [X^1, X^2, \cdots, X^T]$ and $\hat{\mathbf{X}} = [\hat{X}^1, \hat{X}^2, \cdots, \hat{X}^T]$, we define $c(\mathbf{X}, \hat{\mathbf{X}}) = \sum_{t=1}^{T} \|X^t - \hat{X}^t\|_2^2$. With conditional independence assumptions in Equation 1, we express $W_c(P_{\mathbf{X}_{1:M},\mathbf{Y}}, P_{\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}}})$ as:

Proposition 1. *For any functions* $G_y : \mathbf{Z_y} \rightarrow \mathbf{F_y}$, $G_{a\{1:M\}} : \mathbf{Z_{a\{1:M\}}} \rightarrow \mathbf{F_{a\{1:M\}}}$, $D : \mathbf{F_y} \rightarrow \hat{\mathbf{Y}}$, *and* $F_{1:M} : \mathbf{F_{a\{1:M\}}}, \mathbf{F_y} \rightarrow \hat{\mathbf{X}}_{1:M}$, *we have* $W_c(P_{\mathbf{X}_{1:M},\mathbf{Y}}, P_{\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}}}) =$

$$\inf_{Q_{\mathbf{Z}}=P_{\mathbf{Z}}} \mathbf{E}_{P_{\mathbf{X}_{1:M},\mathbf{Y}}} \mathbf{E}_{Q(\mathbf{Z}|\mathbf{X}_{1:M},\mathbf{Y})} \left[\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F_i\big(G_{ai}(\mathbf{Z_{ai}}), G_y(\mathbf{Z_y})\big)\Big) + c_Y\Big(\mathbf{Y}, D\big(G_y(\mathbf{Z_y})\big)\Big)\right], \quad (2)$$

where $P_{\mathbf{Z}}$ *is the prior over* $\mathbf{Z} = [\mathbf{Z_y}, \mathbf{Z_{a\{1,M\}}}]$ *and* $Q_{\mathbf{Z}}$ *is the aggregated posterior of the proposed approximate inference distribution* $Q(\mathbf{Z}|\mathbf{X}_{1:M}, \mathbf{Y})$.

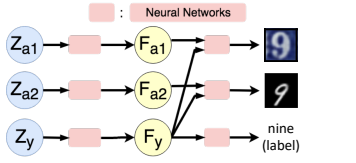
Proof: The proof is adapted from Tolstikhin *et al.* [60]. The two differences are: (1) we show that $P(\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}}|\mathbf{Z} = z)$ are Dirac for all $z \in \mathcal{Z}$, and (2) we use the fact that $c((\mathbf{X}_{1:M}, \mathbf{Y}), (\hat{\mathbf{X}}_{1:M}, \hat{\mathbf{Y}})) = \sum_{i=1}^{M} c_{X\,i}(\mathbf{X}_i, \hat{\mathbf{X}}_i) + c_Y(\mathbf{Y}, \hat{\mathbf{Y}})$. Please refer to the supplementary material for proof details. ■

The constraint on $Q_{\mathbf{Z}} = P_{\mathbf{Z}}$ in Proposition 1 is hard to satisfy. To obtain a numerical solution, we first relax the constraint by performing a generalized mean field assumption on Q according to the conditional independence as shown in the inference network of Figure 1 (b):

$$Q(\mathbf{Z}|\mathbf{X}_{1:M}, \mathbf{Y}) \coloneqq Q(\mathbf{Z}|\mathbf{X}_{1:M}) \coloneqq Q(\mathbf{Z_y}|\mathbf{X}_{1:M}) \prod_{i=1}^{M} Q(\mathbf{Z_{a_i}}|\mathbf{X}_i). \quad (3)$$

The intuition here is based on our design that $\mathbf{Z_y}$ generates the multimodal discriminative factor $\mathbf{F_y}$ and $\mathbf{Z_{a\{1:M\}}}$ generate modality-specific generative factors $\mathbf{F_{a\{1:M\}}}$. Therefore, the inference for $\mathbf{Z_y}$ should depend on all modalities $\mathbf{X}_{1:M}$ and the inference for $\mathbf{Z_{a_i}}$ should depend only on the specific modality $\mathbf{X}_i$. Following this assumption, we define $\mathcal{Q}$ as a nonparametric set of all encoders that fulfill the factorization in Equation 3. A penalty term is added into our objective to find the $Q(\mathbf{Z}|\cdot) \in \mathcal{Q}$ that is the closest to prior $P_{\mathbf{Z}}$, thereby approximately enforcing the constraint $Q_{\mathbf{Z}} = P_{\mathbf{Z}}$:

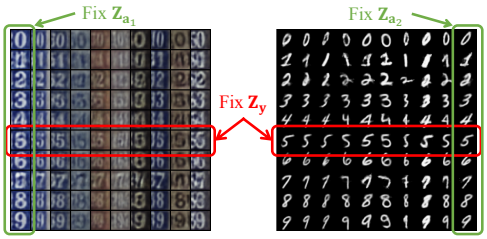
$$\min_{F, G_{a\{1:M\}}, G_y, D} \inf_{Q(\mathbf{Z}|\cdot)\in\mathcal{Q}} \mathbf{E}_{P_{\mathbf{X}_{1:M},\mathbf{Y}}} \mathbf{E}_{Q(\mathbf{Z_{a1}}|\mathbf{X}_1)} \cdots \mathbf{E}_{Q(\mathbf{Z_{aM}}|\mathbf{X}_M)} \mathbf{E}_{Q(\mathbf{Z_y}|\mathbf{X}_{1:M})}$$
$$\left[\sum_{i=1}^{M} c_{X_i}\Big(\mathbf{X}_i, F\big(G_{ai}(\mathbf{Z_{ai}}), G_y(\mathbf{Z_y})\big)\Big) + c_Y\Big(\mathbf{Y}, D\big(G_y(\mathbf{Z_y})\big)\Big)\right] + \lambda \mathcal{MMD}(Q_{\mathbf{Z}}, P_{\mathbf{Z}}), \quad (4)$$



(a)

Method	UM(SVHN)	UM(MNIST)	MM	MFM
Acc.	91.84	99.01	99.20	**99.36**

(b)



(c)

Figure 2: (a) MFM generative network for multimodal image dataset SVHN+MNIST, (b) unimodal and multimodal classification accuracies, and (c) conditional generation for SVHN and MNIST digits. MFM shows improved capabilities in digit prediction as well as flexible generation of both images based on labels and styles.

where λ is a hyper-parameter and $\mathcal{MMD}$ is the Maximum Mean Discrepancy [18] as a divergence measure between $Q_{\mathbf{Z}}$ and $P_{\mathbf{Z}}$. The prior $P_{\mathbf{Z}}$ is chosen as a centered isotropic Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$, so that it implicitly enforces independence between the latent variables $\mathbf{Z} = [\mathbf{Z_y}, \mathbf{Z_{a\{1,M\}}}]$ [21, 26, 51].

Equation 4 represents our hybrid generative-discriminative optimization objective over multimodal data: the first loss term $\sum_{i=1}^{M} c_{X_i}(\mathbf{X}_i, F(G_{ai}(\mathbf{Z_{ai}}), G_y(\mathbf{Z_y})))$ is the generative objective based on reconstruction of multimodal data and the second term $c_Y(\mathbf{Y}, D(G_y(\mathbf{Z_y})))$ is the discriminative objective. In practice we compute the expectations in Equation 4 using empirical estimates over the training data. The neural architecture of MFM is illustrated in Figure 1(c).

2.3 Surrogate Inference for Missing Modalities

A key challenge in multimodal learning involves dealing with missing modalities. A good multimodal model should be able to infer the missing modality conditioned on the observed modalities and perform predictions based only on the observed modalities. To achieve this objective, the inference process of MFM can be easily adapted using a surrogate inference network to reconstruct the missing modality given the observed modalities. Formally, let Φ denote the surrogate inference network. The generation of missing modality $\hat{\mathbf{X}}_1$ given the observed modalities $\mathbf{X}_{2:M}$ can be formulated as

$$\begin{aligned} \Phi^* &= \underset{\Phi}{\operatorname{argmin}} \mathbf{E}_{P_{\mathbf{X}_{2:M}, \hat{\mathbf{X}}_1}} \Big(-\log P_\Phi(\hat{\mathbf{X}}_1 | \mathbf{X}_{2:M}) \Big) \\ \text{with } P_\Phi(\hat{\mathbf{X}}_1 | \mathbf{X}_{2:M}) &:= \int P(\hat{\mathbf{X}}_1 | \mathbf{Z_{a1}}, \mathbf{Z_y}) Q_\Phi(\mathbf{Z_{a1}} | \mathbf{X}_{2:M}) Q_\Phi(\mathbf{Z_y} | \mathbf{X}_{2:M}) d\mathbf{Z_{a1}} d\mathbf{Z_y}. \end{aligned} \tag{5}$$

Similar to Section 2.2, we use deterministic mappings in $Q_\Phi(\cdot|\cdot)$ and $Q_\Phi(\mathbf{Z_y}|\cdot)$ is also used for prediction $P_\Phi(\hat{\mathbf{Y}}|\mathbf{X}_{2:M}) := \int P(\hat{\mathbf{Y}}|\mathbf{Z_y}) Q_\Phi(\mathbf{Z_y}|\mathbf{X}_{2:M}) d\mathbf{Z_y}$. Equation 5 suggests that in the presence of missing modalities, we only need to infer the latent codes rather than the entire modality.

2.4 Encoder and Decoder Design

We now discuss the implementation choices for the MFM neural architecture in Figure 1(c). The encoder $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$ can be parametrized by any model that performs multimodal fusion [36, 67]. For multimodal image datasets, we adopt Convolutional Neural Networks (CNNs) and Fully-Connected Neural Networks (FCNNs) with late fusion [40] as our encoder $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$. The remaining functions in MFM are also parametrized by CNNs and FCNNs. For multimodal time series datasets, we choose the Memory Fusion Network (MFN) [68] as our multimodal encoder $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$. We use Long Short-term Memory (LSTM) networks [22] for functions $Q(\mathbf{Z_{a\{1:M\}}}|\mathbf{X}_{1:M})$, decoder LSTM networks [10] for functions $F_{1:M}$, and FCNNs for functions G_y, $G_{a\{1:M\}}$ and D. Details are provided in the appendix.

3 Experiments

In order to show that MFM learns multimodal representations that are discriminative, generative and interpretable, we design the following experiments. We begin with a multimodal synthetic image dataset that allows us to examine whether MFM displays discriminative and generative capabilities from factorized latent variables. Utilizing image datasets allows us to clearly visualize the generative capabilities of MFM. We then transition to six more challenging real-world multimodal video datasets to 1) rigorously evaluate the discriminative capabilities of MFM in comparison with existing baselines,

Table 1: Results on CMU-MOSI, ICT-MMMO, YouTube, and MOUD. SOTA1 and SOTA2 refer to the previous best and second best state-of-the-art respectively, and Δ_{SOTA} shows improvement over SOTA1. Symbols depict the baseline giving the result: $\#$ *MFN*, ‡ *MARN*, $*$ *TFN*, † *BC-LSTM*, $\diamond$ *MV-LSTM*, § *EF-LSTM*, $\flat$ *DF*, $\heartsuit$ *SVM*, • *RF*. For detailed results for all models, please refer to the appendix.

Dataset	POM Personality Traits															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	r															
SOTA2	0.359†	0.425†	0.166‡	0.235‡	0.358†	0.417†	0.450†	0.378‡	0.295$^{\diamond}$	0.237$^{\diamond}$	0.215‡	0.238$^{\diamond}$	0.363†	0.258$^{\diamond}$	0.344†	0.319†
SOTA1	0.395$^{\#}$	0.428$^{\#}$	0.193$^{\#}$	0.313$^{\#}$	0.367$^{\#}$	0.431$^{\#}$	0.452$^{\#}$	0.395$^{\#}$	0.333$^{\#}$	**0.296**$^{\#}$	0.255$^{\#}$	0.259$^{\#}$	0.381$^{\#}$	0.318$^{\#}$	**0.377**$^{\#}$	0.386$^{\#}$
MFM	**0.431**	**0.450**	**0.197**	**0.411**	**0.380**	**0.448**	**0.467**	**0.452**	**0.368**	0.212	**0.309**	**0.333**	**0.404**	**0.333**	0.334	**0.408**
Δ_{SOTA}	↑ 0.036	↑ 0.022	↑ 0.004	↑ 0.097	↑ 0.013	↑ 0.017	↑ 0.015	↑ 0.057	↑ 0.035	–	↑ 0.054	↑ 0.074	↑ 0.023	↑ 0.015	–	↑ 0.022

Dataset	CMU-MOSI					ICT-MMMO		YouTube		MOUD	
Task	Sentiment					Sentiment		Sentiment		Sentiment	
Metric	Acc_7	Acc_2	F1	MAE	r	Acc_2	F1	Acc_3	F1	Acc_2	F1
SOTA2	34.1$^{\#}$	77.1‡	77.0‡	0.968‡	0.625‡	72.5*	72.6*	48.3‡	45.1†	81.1$^{\#}$	80.4$^{\#}$
SOTA1	34.7‡	77.4$^{\#}$	77.3$^{\#}$	0.965$^{\#}$	0.632$^{\#}$	73.8$^{\#}$	73.1$^{\#}$	51.7$^{\#}$	51.6$^{\#}$	81.1‡	81.2‡
MFM	**36.2**	**78.1**	**78.1**	**0.951**	**0.662**	**81.3**	**79.2**	**53.3**	**52.4**	**82.1**	**81.7**
Δ_{SOTA}	↑ 1.5	↑ 0.7	↑ 0.8	↓ 0.014	↑ 0.030	↑ 7.5	↑ 6.1	↑ 1.6	↑ 0.8	↑ 1.0	↑ 0.5

Dataset	IEMOCAP Emotions											
Task	Happy		Sad		Angry		Frustrated		Excited		Neutral	
Metric	Acc_2	F1	Acc_2	F1	Acc_2	F1	Acc_2	F1	Acc_2	F1	Acc_2	F1
SOTA2	86.7‡	84.2§	83.4*	81.7†	85.1$^{\diamond}$	84.5§	79.5‡	76.6‡	89.6‡	86.3$^{\#}$	68.8§	67.1§
SOTA1	90.1$^{\#}$	85.3$^{\#}$	85.8$^{\#}$	82.8*	87.0$^{\#}$	86.0$^{\#}$	80.3$^{\#}$	**76.8**$^{\#}$	89.8$^{\#}$	**87.1**‡	71.8$^{\#}$	**68.5**§
MFM	**90.2**	**85.8**	**88.4**	**86.1**	**87.5**	**86.7**	**80.4**	74.5	**90.0**	**87.1**	**72.1**	68.1
Δ_{SOTA}	↑ 0.1	↑ 0.5	↑ 2.6	↑ 3.3	↑ 0.5	↑ 0.7	↑ 0.1	–	↑ 0.2	–	↑ 0.3	–

2) analyze the importance of each design component through ablation studies, 3) assess the robustness of MFM's modality reconstruction and prediction capabilities to missing modalities, and 4) interpret the learned representations using information-based and gradient-based methods to understand the contributions of individual factors towards multimodal prediction and generation.

3.1 Multimodal Synthetic Image Dataset

In this section, we study MFM on a synthetic image dataset that considers SVHN [38] and MNIST [30] as the two modalities. SVHN and MNIST are images with different styles but the same labels (digits $0 \sim 9$). We randomly pair $100,000$ SVHN and MNIST images that have the same label, creating a multimodal dataset which we call SVHN+MNIST. $80,000$ pairs are used for training and the rest for testing. To justify that MFM is able to learn improved multimodal representations, we show both classification and generation results on SVHN+MNIST in Figure 2.

Prediction: We perform experiments on both unimodal and multimodal classification tasks. UM denotes a unimodal baseline that performs prediction given only one modality as input and MM denotes a multimodal discriminative baseline that performs prediction given both images [40]. We compare the results for UM(SVHN), UM(MNIST), MM and MFM on SVHN+MNIST in Figure 2(b). We achieve better classification performance from unimodal to multimodal which is not surprising since more information is given. More importantly, MFM outperforms MM, which suggests that MFM learns improved factorized representations for discriminative tasks.

Generation: We generate images using the MFM generative network (Figure 2(a)). We fix one variable out of $\mathbf{Z} = [\mathbf{Z_{a1}}, \mathbf{Z_{a2}}, \text{and } \mathbf{Z_y}]$ and randomly sample the other two variables from prior $P_{\mathbf{Z}}$. From Figure 2(c), we observe that MFM shows flexible generation of SVHN and MNIST images based on labels and styles. This suggests that MFM is able to factorize multimodal representations into multimodal discriminative factors (labels) and modality-specific generative factors (styles).

3.2 Multimodal Time Series Datasets

In this section, we transition to more challenging multimodal time series datasets. All the datasets consist of monologue videos. Features are extracted from the language (GloVe word embeddings [42]), visual (Facet [24]), and acoustic (COVAREP [12]) modalities. For a detailed description of feature extraction, please refer to the appendix.

Model	Multimodal Disc. Factor	Hybrid Gen.-Disc. Objective	Factorized Gen.-Disc. Factors	Mod.-Spec. Gen. Factors	CMU-MOSI							
					$\hat{\mathbf{X}}$. Reconstruction			$\hat{\mathbf{Y}}$ Prediction				
					MSE (ℓ)	MSE (a)	MSE (v)	Acc_7	Acc_2	F1	MAE	r
$\mathbf{M_A}$	no	no	-	-	-	-	-	33.2	75.2	75.2	1.020	0.616
$\mathbf{M_B}$	yes	no	-	-	-	-	-	34.1	77.4	77.3	0.965	0.632
$\mathbf{M_C}$	no	yes	no	-	0.0413	0.0509	0.0220	34.8	75.9	76.0	0.979	0.640
$\mathbf{M_D}$	yes	yes	no	-	0.0413	0.0486	0.0223	35.0	77.4	77.2	0.960	0.649
$\mathbf{M_E}$	yes	yes	yes	no	0.0397	0.0452	0.0211	35.9	77.3	77.2	0.956	0.661
MFM	yes	yes	yes	yes	**0.0391**	**0.0384**	**0.0183**	**36.2**	**78.1**	**78.1**	**0.951**	**0.662**

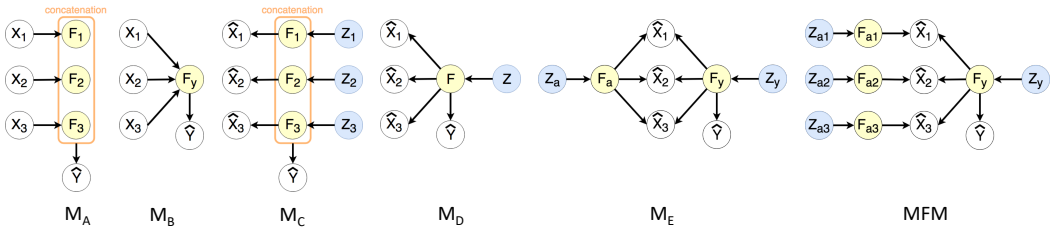


Figure 3: Models used in the ablation studies of MFM. Each model removes a design component from our model. Modality reconstruction and sentiment prediction results are reported on CMU-MOSI with best results in bold. Factorizing multimodal representations into multimodal discriminative factors and modality-specific generative factors are crucial for improved performance.

We consider the following six datasets across three domains: 1) Multimodal Personality Trait Recognition: **POM** [41] contains 903 movie review videos annotated for the following personality traits: confident (con), passionate (pas), voice pleasant (voi), dominant (dom), credible (cre), vivid (viv), expertise (exp), entertaining (ent), reserved (res), trusting (tru), relaxed (rel), outgoing (out), thorough (tho), nervous (ner), persuasive (per) and humorous (hum). The short form is indicated in parenthesis. 2) Multimodal Sentiment Analysis: **CMU-MOSI** [70] is a collection of 2199 monologue opinion video clips annotated with sentiment. **ICT-MMMO** [63] consists of 340 online social review videos annotated for sentiment. **YouTube** [36] contains 269 product review and opinion video segments from YouTube each annotated for sentiment. **MOUD** [43] consists of 79 product review videos in Spanish. Each video consists of multiple segments labeled as either positive, negative or neutral sentiment. 3) Multimodal Emotion Recognition: **IEMOCAP** [4] consists of 302 videos of recorded dyadic dialogues. The videos are divided into multiple segments each annotated for the presence of 6 discrete emotions (happy, sad, angry, frustrated, excited and neutral), resulting in a total of 7318 segments in the dataset. We report results using the following metrics: Acc_C = multiclass accuracy across C classes, F1 = F1 score, MAE = Mean Absolute Error, r = Pearson's correlation.

Prediction: We first compare the performance of MFM with existing multimodal prediction methods. For a detailed description of the baselines, please refer to the appendix. From Table 1, we first observe that the best performing baseline results are achieved by different models across different datasets (most notably MFN, MARN, and TFN). On the other hand, MFM consistently achieves state-of-the-art or competitive results for all six multimodal datasets. We believe that the multimodal discriminative factor $\mathbf{F_y}$ in MFM has successfully learned more meaningful representations by distilling discriminative features. This highlights the benefit of learning factorized multimodal representations towards discriminative tasks.

Furthermore, MFM is *model-agnostic* and can be applied to other multimodal encoders $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$. We perform experiments to show consistent improvements in discriminative performance for several choices of the encoder: EF-LSTM [36] and TFN [67]. For Acc_2 on CMU-MOSI, our factorization framework improves the performance of EF-LSTM from 74.3 to **75.2** and TFN from 74.6 to **75.5**.

Ablation Study: In Figure 3, we present the models $\mathbf{M}_{\{\mathbf{A},\mathbf{B},\mathbf{C},\mathbf{D},\mathbf{E}\}}$ used for ablation studies. These models are designed to analyze the effects of using a multimodal discriminative factor, a hybrid generative-discriminative objective, factorized generative-discriminative factors and modality-specific generative factors towards both modality reconstruction and label prediction. The simplest variant is $\mathbf{M_A}$ which represents a purely discriminative model without a joint multimodal discriminative factor (i.e. early fusion [36]). $\mathbf{M_B}$ models a joint multimodal discriminative factor which incorporates more general multimodal fusion encoders [68]. $\mathbf{M_C}$ extends $\mathbf{M_A}$ by optimizing a hybrid generative-discriminative objective over modality-specific factors. $\mathbf{M_D}$ extends $\mathbf{M_B}$ by optimizing a hybrid generative-discriminative objective over a joint multimodal factor (resembling [57]). $\mathbf{M_E}$ factorizes

Table 2: The effect of missing modalities on multimodal data reconstruction and sentiment prediction on CMU-MOSI. MFM with surrogate inference is able to better handle missing modalities during test time as compared to the purely generative (Seq2Seq) or purely discriminative baselines.

Task	$\hat{\mathbf{X}}$. Reconstruction			$\hat{\mathbf{Y}}$ Prediction				
Metric	MSE (ℓ)	MSE (a)	MSE (v)	Acc_7	Acc_2	F1	MAE	r
Purely Generative and Discriminative Baselines								
ℓ(anguage) missing	0.0411	-	-	19.4	59.6	59.7	1.386	0.225
a(udio) missing	-	0.0533	-	34.0	73.5	73.4	1.024	0.615
v(isual) missing	-	-	0.0220	33.7	75.4	75.4	0.996	0.634
Multimodal Factorization Model (MFM)								
ℓ(anguage) missing	0.0403	-	-	21.7	62.0	61.7	1.313	0.236
a(udio) missing	-	0.0468	-	35.4	74.3	74.3	1.011	0.603
v(isual) missing	-	-	0.0215	35.0	76.4	76.3	0.990	0.635
all present	**0.0391**	**0.0384**	**0.0182**	**36.2**	**78.1**	**78.1**	**0.951**	**0.662**

the representation into separate generative and discriminative factors. Finally, MFM is obtained from $\mathbf{M_E}$ by using modality-specific generative factors instead of a joint multimodal generative factor.

From the table in Figure 3, we observe the following general trends. For sentiment prediction, using 1) a multimodal discriminative factor outperforms modality-specific discriminative factors ($\mathbf{M_D} > \mathbf{M_C}$, $\mathbf{M_B} > \mathbf{M_A}$), and 2) adding generative capabilities to the model improves performance ($\mathbf{M_C} > \mathbf{M_A}$, $\mathbf{M_E} > \mathbf{M_B}$). For both sentiment prediction and modality reconstruction, 3) factorizing into separate generative and discriminative factors improves performance ($\mathbf{M_E} > \mathbf{M_D}$), and 4) using modality-specific generative factors outperforms multimodal generative factors (MFM > $\mathbf{M_E}$). These observations support our design decisions of factorizing multimodal representations into multimodal discriminative factors and modality-specific generative factors.

Missing Modalities: We now evaluate the performance of MFM in the presence of missing modalities using the surrogate inference model as described in Subsection 2.3. We compare with two baselines: 1) a purely generative Seq2Seq model [10] Φ_G from observed modalities to missing modalities by optimizing $\mathbf{E}_{P_{\mathbf{X}_{1:M}}}\left(-\log P_{\Phi_D}(\mathbf{X}_1|\mathbf{X}_{2:M})\right)$, and 2) a purely discriminative model Φ_D from observed modalities to the label by optimizing $\mathbf{E}_{P_{\mathbf{X}_{2:M},\mathbf{Y}}}\left(-\log P_{\Phi_G}(\mathbf{Y}|\mathbf{X}_{2:M})\right)$. Both models are modified from MFM by using only the two observed modalities as input and not explicitly accounting for missing modalities. We compare the reconstruction error of each modality (language, visual and acoustic) as well as the performance on sentiment prediction.

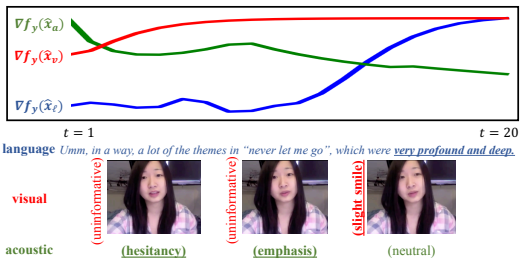
Table 2 shows that MFM with missing modalities outperforms the generative (Φ_G) or discriminative baselines (Φ_D) in terms of modality reconstruction and sentiment prediction. Additionally, MFM with missing modalities performs close to MFM with all modalities observed. This fact indicates that MFM can learn representations that are relatively robust to missing modalities. In addition, discriminative performance is most affected when the language modality is missing, which is consistent with prior work which indicates that language is most informative in human multimodal language [67]. On the other hand, sentiment prediction is more robust to missing acoustic and visual features. Finally, we observe that reconstructing the low-level acoustic and visual features is easier as compared to the high-dimensional language features that contain high-level semantic meaning.

Interpretation of Multimodal Representations: We devise two methods to study how individual factors in MFM influence the dynamics of multimodal prediction and generation. These interpretation methods represent both overall trends and fine-grained analysis that could be useful towards deeper understandings of multimodal representation learning. For more details, please refer to the appendix.

Firstly, an information-based interpretation method is chosen to summarize the contribution of each modality towards the multimodal representations. Since $\mathbf{F_y}$ is a common cause of $\hat{\mathbf{X}}_{1:M}$, we can compare $\text{MI}(\mathbf{F_y}, \hat{\mathbf{X}}_1), \cdots, \text{MI}(\mathbf{F_y}, \hat{\mathbf{X}}_M)$, where $\text{MI}(\cdot,\cdot)$ denotes the mutual information measure between $\mathbf{F_y}$ and generated modality $\hat{\mathbf{X}}_i$. Higher $\text{MI}(\mathbf{F_y}, \hat{\mathbf{X}}_i)$ indicates greater contribution from $\mathbf{F_y}$ to $\hat{\mathbf{X}}_i$. Figure 4 reports the ratios $r_i = \text{MI}(\mathbf{F_y}, \hat{\mathbf{X}}_i)/\text{MI}(\mathbf{F}_{\mathbf{a}i}, \hat{\mathbf{X}}_i)$ which measure a normalized version of the mutual information between $\mathbf{F}_{\mathbf{a}i}$ and $\hat{\mathbf{X}}_i$. We observe that on CMU-MOSI, the language modality is most informative towards sentiment prediction, followed by the acoustic modality. We believe that this result represents a prior over the expression of sentiment in human multimodal language and is closely related to the connections between language and speech [27].

Ratio	r_ℓ	r_v	r_a
CMU-MOSI	0.307	0.030	0.107

Figure 4: Analyzing the multimodal representations learnt in MFM via information-based (entire dataset) and gradient-based interpretation methods (single video) on CMU-MOSI.



Secondly, a gradient-based interpretation method to used analyze the contribution of each modality for every time step in multimodal time series data. We measure the gradient of the generated modality with respect to the target factors (e.g., $\mathbf{F_y}$). Let $\{x_1, x_2, \cdots, x_M\}$ denote multimodal time series data where x_i represents modality i, and $\hat{x}_i = [\hat{x}_i^1, \cdots, \hat{x}_i^t, \cdots, \hat{x}_i^T]$ denote generated modality i across time steps $t \in [1, T]$. The gradient $\nabla_{f_y}(\hat{x}_i)$ measures the extent to which changes in factor $f_y \sim P(\mathbf{F_y}|\mathbf{X}_{1:M} = x_{1:M})$ influences the generation of sequence $\hat{x}_i$. Figure 4 plots $\nabla_{f_y}(\hat{x}_i)$ for a video in CMU-MOSI. We observe that multimodal communicative behaviors that are indicative of speaker sentiment such as positive words (e.g. "very profound and deep") and informative acoustic features (e.g. hesitant and emphasized tone of voice) indeed correspond to increases in $\nabla_{f_y}(\hat{x}_i)$.

4 Related Work

The two main pillars of research in multimodal representation learning have considered the discriminative and generative objectives individually. Discriminative representation learning [32, 6, 5, 16, 53, 61] models the conditional distribution $P(\mathbf{Y}|\mathbf{X}_{1:M})$. Since these approaches are not concerned with modeling $P(\mathbf{X}_{1:M})$ explicitly, they use parameters more efficiently to model $P(\mathbf{Y}|\mathbf{X}_{1:M})$. For instance, recent works learn visual representations that are maximally dependent with linguistic attributes for improving one-shot image recognition [62] or introduce tensor product mechanisms to model interactions between the language, visual and acoustic modalities [34, 67]. On the other hand, generative representation learning captures the interactions between modalities by modeling the joint distribution $P(\mathbf{X}_1, \cdots, \mathbf{X}_M)$ using either undirected graphical models [57], directed graphical models [59], or neural networks [54]. Some generative approaches compress multimodal data into lower-dimensional feature vectors which can be used for discriminative tasks [45, 39]. To unify the advantages of both approaches, MFM factorizes multimodal representations into generative and discriminative components and optimizes for a joint objective.

Factorized representation learning resembles learning disentangled data representations which have been shown to improve the performance on many tasks [28, 29, 21, 2]. Several methods involve specifying a fixed set of latent attributes that individually control particular variations of data and performing supervised training [8, 25, 65, 50, 72], assuming an isotropic Gaussian prior over latent variables to learn disentangled generative representations [26, 51] and learning latent variables in charge of specific variations in the data by maximizing the mutual information between a subset of latent variables and the data [7]. However, these methods study factorization of a single modality. MFM factorizes multimodal representations and demonstrates the importance of modality-specific and multimodal factors towards generation and prediction. A concurrent and parallel work that factorizes latent factors in multimodal data was proposed by [23]. They differ from us in the graphical model design, discriminative objective, prior matching criterion, and scale of experiments. We provide a detailed comparison with their model in the appendix.

5 Conclusion

In this paper, we proposed the Multimodal Factorization Model (MFM) for multimodal representation learning. MFM factorizes the multimodal representations into two sets of independent factors: *multimodal discriminative* factors and *modality-specific generative* factors. The multimodal discriminative factor achieves state-of-the-art or competitive results on six multimodal datasets. The modality-specific generative factors allow us to generate data based on factorized variables, account for missing modalities, and have a deeper understanding of the interactions involved in multimodal learning. Our future work will explore extensions of MFM for video generation, semi-supervised learning, and unsupervised learning. We believe that MFM sheds light on the advantages of learning factorizing multimodal representations and potentially opens up new horizons for multimodal machine learning.

References

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *arXiv preprint arXiv:1705.09406*, 2017.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8), August 2013.

[3] Leo Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, October 2001.

[4] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation*, 2008.

[5] Devendra Singh Chaplot, Kanthashree Mysore Sathyendra, Rama Kumar Pasumarthi, Dheeraj Rajagopal, and Ruslan Salakhutdinov. Gated-attention architectures for task-oriented language grounding. *arXiv preprint arXiv:1706.07230*, 2017.

[6] Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrušaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI, 2017.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2172–2180, 2016.

[8] Brian Cheung, Jesse A Livezey, Arjun K Bansal, and Bruno A Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

[9] KyungHyun Cho, Aaron C. Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *CoRR*, abs/1507.01053, 2015.

[10] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*. ACL, 2014.

[11] Marco Cuturi, Jean-Philippe Vert, Oystein Birkenes, and Tomoko Matsui. A kernel for time series based on global alignments. In *Acoustics, Speech and Signal Processing, ICASSP 2007.*, 2007.

[12] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarepa collaborative voice analysis repository for speech technologies. In *ICASSP*. IEEE, 2014.

[13] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[14] Paul Ekman, Wallace V Freisen, and Sonia Ancoli. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6):1125, 1980.

[15] C. A. Frantzidis, C. Bratsas, M. A. Klados, E. Konstantinidis, C. D. Lithari, A. B. Vivas, C. L. Papadelis, E. Kaldoudi, C. Pappas, and P. D. Bamidis. On the classification of emotional biosignals evoked while viewing affective pictures: An integrated data-mining-based approach for healthcare applications. *IEEE Transactions on Information Technology in Biomedicine*, March 2010.

[16] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[17] A. Graves, A. r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013.

[18] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[19] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[20] Sylvain Guimond and Wael Massrieh. Intricate correlation between body posture, personality trait and incidence of body pain: A cross-referential study report. *PLOS ONE*, 2012.

[21] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[22] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[23] Wei-Ning Hsu and James Glass. Disentangling by partitioning: A representation learning framework for multimodal sensory data. *arXiv preprint arXiv:1805.11264*, 2018.

[24] iMotions. Facial expression analysis, 2017.

[25] Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[27] Patricia K. Kuhl. A new view of language acquisition. *Proceedings of the National Academy of Sciences*, 2000.

[28] Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in Neural Information Processing Systems*, 2015.

[29] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, 2017.

[30] Yann Lecun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324, 1998.

[31] Sergey Levine, Christian Theobalt, and Vladlen Koltun. Real-time prosody-driven synthesis of body language. *ACM Trans. Graph.*, 28(5):172:1–172:10, December 2009.

[32] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Multimodal local-global ranking fusion for emotion recognition. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, ICMI, 2018.

[33] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao. A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica*, 4(4):668–676, 2017.

[34] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *ACL*, 2018.

[35] Gelareh Mohammadi, Alessandro Vinciarelli, and Marcello Mortillaro. The voice of personality: Mapping nonverbal vocal behavior into trait attributions. In *Proceedings of ACM Multimedia Workshop on Social Signal Processing*, 0 2010.

[36] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*, pages 169–176. ACM, 2011.

[37] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *CVPR*. IEEE, 2007.

[38] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. 2011.

[39] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *(ICML-11)*, 2011.

[40] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI 2016. ACM, 2016.

[41] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Philippe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. ICMI '14, 2014.

[42] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[43] Veronica Perez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. Utterance-Level Multimodal Sentiment Analysis. In *Association for Computational Linguistics (ACL)*, August 2013.

[44] Sintija Petrovica, Alla Anohina-Naumeca, and HazÄ±m Kemal Ekenel. Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science*, 2017.

[45] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*. ACL, 2018.

[46] Johannes Pittermann, Angela Pittermann, and Wolfgang Minker. Emotion recognition and adaptation in spoken dialogue systems. *International Journal of Speech Technology*, 13(1):49–60, Mar 2010.

[47] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *ACL*, 2017.

[48] Ariadna Quattoni, Sybor Wang, Louis-Philippe Morency, Michael Collins, and Trevor Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2007.

[49] Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis, and Goecke Roland. Extending long short-term memory for multi-view structured learning. In *ECCV*, 2016.

[50] Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, 2014.

[51] Paul K Rubenstein, Bernhard Schoelkopf, and Ilya Tolstikhin. On the latent space of wasserstein auto-encoders. *arXiv preprint arXiv:1802.03761*, 2018.

[52] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11), November 1997.

[53] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[54] Kihyuk Sohn, Wenling Shang, and Honglak Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014.

[55] Yale Song, Louis-Philippe Morency, and Randall Davis. Multi-view latent variable discriminative models for action recognition. In *CVPR*, 2012.

[56] Yale Song, Louis-Philippe Morency, and Randall Davis. Action recognition by hierarchical sequence summarization. In *CVPR*, 2013.

[57] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012.

[58] Masashi Sugiyama and Makoto Yamada. On kernel parameter selection in hilbert-schmidt independence criterion. *IEICE TRANSACTIONS on Information and Systems*, 2012.

[59] Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. Joint multimodal learning with deep generative models. *arXiv preprint arXiv:1611.01891*, 2016.

[60] Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*, 2017.

[61] Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. Learning robust visual-semantic embeddings. *arXiv preprint arXiv:1703.05908*, 2017.

[62] Yao-Hung Hubert Tsai and Ruslan Salakhutdinov. Improving one-shot learning through fusing side information. *arXiv preprint arXiv:1710.08347*, 2017.

[63] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. Youtube movie reviews: Sentiment analysis in an audio-visual context. *IEEE Intelligent Systems*, 28(3):46–53, 2013.

[64] Denny Wu, Yixiu Zhao, Yao-Hung Hubert Tsai, Makoto Yamada, and Ruslan Salakhutdinov. " dependency bottleneck" in auto-encoding architectures: an empirical study. *arXiv preprint arXiv:1802.05408*, 2018.

[65] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. In *NIPS*, 2015.

[66] Jiahong Yuan and Mark Liberman. Speaker identification on the scotus corpus. *Journal of the Acoustical Society of America*, 123(5):3878, 2008.

[67] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, 2017.

[68] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. *AAAI*, 2018.

[69] Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. *AAAI*, 2018.

[70] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 2016.

[71] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders. *arXiv preprint arXiv:1706.02262*, 2017.

[72] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Multi-view perceptron: a deep model for learning face identity and view representations. In *NIPS*, 2014.

A Proof of Proposition 1

To simplify the proof, we first prove it for the unimodal case by considering the Wasserstein distance between $P_{\mathbf{X},\mathbf{Y}}$ and $P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}$.

A.1 Unimodal Joint-Distribution Wasserstein Distance

Proposition 2. *For any functions $G_y : \mathbf{Z_y} \to \mathbf{F_y}$, $G_a : \mathbf{Z_a} \to \mathbf{F_a}$, $D : \mathbf{F_y} \to \hat{\mathbf{Y}}$, and $F : \mathbf{F_a}, \mathbf{F_y} \to \hat{\mathbf{X}}$, we have*

$$W_c(P_{\mathbf{X},\mathbf{Y}}, P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}) = \inf_{Q_{\mathbf{Z}}=P_{\mathbf{Z}}} \mathbf{E}_{P_{\mathbf{X},\mathbf{Y}}} \mathbf{E}_{Q(\mathbf{Z}|\mathbf{X})} \Bigg[c_X\Big(\mathbf{X}, F\big(G_a(\mathbf{Z_a}), G_y(\mathbf{Z_y})\big)\Big) + c_Y\Big(\mathbf{Y}, D\big(G_y(\mathbf{Z_y})\big)\Big) \Bigg], \tag{6}$$

where W_c is the Wasserstein distance under cost function c_X and c_Y, $P_{\mathbf{Z}}$ is the prior over $\mathbf{Z} = [\mathbf{Z_a}, \mathbf{Z_y}]$ and $Q_{\mathbf{Z}}$ is the aggregated posterior of the proposed inference distribution $Q(\mathbf{Z}|\mathbf{X})$.

Proof: See the following.

To begin the proof, we abuse some notations as follows.

By definition, the Wasserstein distance under cost function c between $P_{\mathbf{X},\mathbf{Y}}$ and $P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}$ is

$$W_c(P_{\mathbf{X},\mathbf{Y}}, P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}) := \inf_{\Gamma \in \mathcal{P}\Big((\mathbf{X},\mathbf{Y}) \sim P_{\mathbf{X},\mathbf{Y}}, (\hat{\mathbf{X}},\hat{\mathbf{Y}}) \sim P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}\Big)} \mathbf{E}_{\big((\mathbf{X},\mathbf{Y}),(\hat{\mathbf{X}},\hat{\mathbf{Y}})\big)\sim\Gamma}\Big[c\Big((\mathbf{X},\mathbf{Y}),(\hat{\mathbf{X}},\hat{\mathbf{Y}})\Big)\Big], \tag{7}$$

where $c\Big((\mathbf{X},\mathbf{Y}),(\hat{\mathbf{X}},\hat{\mathbf{Y}})\Big) : (\mathcal{X},\mathcal{Y}) \times (\mathcal{X},\mathcal{Y}) \to \mathcal{R}_+$ is any measurable *cost function*. $\mathcal{P}\Big((\mathbf{X},\mathbf{Y}) \sim P_{\mathbf{X},\mathbf{Y}}, (\hat{\mathbf{X}},\hat{\mathbf{Y}}) \sim P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}\Big)$ is the set of all joint distributions of $\Big((\mathbf{X},\mathbf{Y}),(\hat{\mathbf{X}},\hat{\mathbf{Y}})\Big)$ with marginals $P_{\mathbf{X},\mathbf{Y}}$ and $P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}$, respectively. Note that $c\Big((\mathbf{X},\mathbf{Y}),(\hat{\mathbf{X}},\hat{\mathbf{Y}})\Big) = c_X\Big(\mathbf{X},\hat{\mathbf{X}}\Big) + c_Y\Big(\mathbf{Y},\hat{\mathbf{Y}}\Big)$.

Next, we denote the set of all joint distributions of $(\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z})$ such that $(\mathbf{X},\mathbf{Y}) \sim P_{\mathbf{X},\mathbf{Y}}$, $(\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}) \sim P_{\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}}$, and $\Big((\mathbf{X},\mathbf{Y}) \perp\!\!\!\perp (\hat{\mathbf{X}},\hat{\mathbf{Y}})|\mathbf{Z}\Big)$ as $\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}}$. $\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}}}$ and $\mathcal{P}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}$ are the sets of the marginals $(\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}})$ and $(\mathbf{X},\mathbf{Y},\mathbf{Z})$ induced by $\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}}$.

We now introduce two Lemmas to help the proof.

Lemma 1. *$P(\hat{\mathbf{X}},\hat{\mathbf{Y}}|\mathbf{Z} = z)$ are Dirac for all $z \in \mathcal{Z}$.*

Proof: First, we have $\hat{\mathbf{X}} = F(G_a(\mathbf{Z_a}), G_y(\mathbf{Z_y}))$ and $\hat{\mathbf{Y}} = D(G_y(\mathbf{Z_y}))$ with $\mathbf{Z} = \{\mathbf{Z_a}, \mathbf{Z_y}\}$. Since the functions F, G_a, G_y, D are all deterministic, then $P(\hat{\mathbf{X}},\hat{\mathbf{Y}}|\mathbf{Z})$ are Dirac measures. □

Lemma 2. *$\mathcal{P}\Big(P_{\mathbf{X},\mathbf{Y}}, P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}\Big) = \mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}}}$ when $P(\hat{\mathbf{X}},\hat{\mathbf{Y}}|\mathbf{Z} = z)$ are Dirac for all $z \in \mathcal{Z}$.*

Proof: When $\hat{\mathbf{X}},\hat{\mathbf{Y}}$ are deterministic functions of $\mathbf{Z}$, for any A in the sigma-algebra induced by $\hat{\mathbf{X}},\hat{\mathbf{Y}}$, we have

$$\mathbf{E}[\mathbb{I}_{[\hat{\mathbf{X}},\hat{\mathbf{Y}}\in A]}|\mathbf{X},\mathbf{Y},\mathbf{Z}] = \mathbf{E}[\mathbb{I}_{[\hat{\mathbf{X}},\hat{\mathbf{Y}}\in A]}|\mathbf{Z}].$$

Therefore, this implies that $(\mathbf{X},\mathbf{Y}) \perp\!\!\!\perp (\hat{\mathbf{X}},\hat{\mathbf{Y}})|\mathbf{Z}$ which concludes the proof. A similar argument is made in Lemma 1 of [60].

□

Now, we use the fact that $\mathcal{P}\Big(P_{\mathbf{X},\mathbf{Y}}, P_{\hat{\mathbf{X}},\hat{\mathbf{Y}}}\Big) = \mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}}}$ (Lemma 1 + Lemma 2), $c\Big((\mathbf{X},\mathbf{Y}),(\hat{\mathbf{X}},\hat{\mathbf{Y}})\Big) = c_X\Big(\mathbf{X},\hat{\mathbf{X}}\Big) + c_Y\Big(\mathbf{Y},\hat{\mathbf{Y}}\Big)$, $\hat{\mathbf{X}} = F\big(G_a(\mathbf{Z_a}), G_y(\mathbf{Z_y})\big)$, and $\hat{\mathbf{Y}} = D\big(G_y(\mathbf{Z_y})\big)$,

Eq. equation 7 becomes

$$
\begin{aligned}
&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}}}} \mathbf{E}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}}\sim P}\Big[c_X\Big(\mathbf{X},\hat{\mathbf{X}}\Big)+c_Y\Big(\mathbf{Y},\hat{\mathbf{Y}}\Big)\Big] \\
=&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}}} \mathbf{E}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}\sim P}\Big[c_X\Big(\mathbf{X},\hat{\mathbf{X}}\Big)+c_Y\Big(\mathbf{Y},\hat{\mathbf{Y}}\Big)\Big] \\
=&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}}} \mathbf{E}_{P_\mathbf{Z}}\mathbf{E}_{P(\mathbf{X},\mathbf{Y}|\mathbf{Z})}\mathbf{E}_{P(\hat{\mathbf{X}},\hat{\mathbf{Y}}|\mathbf{Z})}\Big[c_X\Big(\mathbf{X},\hat{\mathbf{X}}\Big)+c_Y\Big(\mathbf{Y},\hat{\mathbf{Y}}\Big)\Big] \\
=&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\hat{\mathbf{X}},\hat{\mathbf{Y}},\mathbf{Z}}} \mathbf{E}_{P_\mathbf{Z}}\mathbf{E}_{P(\mathbf{X},\mathbf{Y}|\mathbf{Z})}\Big[c_X\Big(\mathbf{X},F\big(G_a(\mathbf{Z_a}),G_y(\mathbf{Z_y})\big)\Big)+c_Y\Big(\mathbf{Y},D\big(G_y(\mathbf{Z_y})\big)\Big)\Big] \\
=&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}} \mathbf{E}_{P_\mathbf{Z}}\mathbf{E}_{P(\mathbf{X},\mathbf{Y}|\mathbf{Z})}\Big[c_X\Big(\mathbf{X},F\big(G_a(\mathbf{Z_a}),G_y(\mathbf{Z_y})\big)\Big)+c_Y\Big(\mathbf{Y},D\big(G_y(\mathbf{Z_y})\big)\Big)\Big] \\
=&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}} \mathbf{E}_{\mathbf{X},\mathbf{Y},\mathbf{Z}\sim P}\Big[c_X\Big(\mathbf{X},F\big(G_a(\mathbf{Z_a}),G_y(\mathbf{Z_y})\big)\Big)+c_Y\Big(\mathbf{Y},D\big(G_y(\mathbf{Z_y})\big)\Big)\Big].
\end{aligned}
\tag{8}
$$

Note that in Eq. equation 8, $\mathcal{P}_{\mathbf{X},\mathbf{Y},\mathbf{Z}} = \mathcal{P}\Big((\mathbf{X},\mathbf{Y})\sim P_{\mathbf{X},\mathbf{Y}},\mathbf{Z}\sim P_\mathbf{Z}\Big)$ and with a proposed $Q(\mathbf{Z}|\mathbf{X})$, we can rewrite Eq. equation 8 as

$$
\begin{aligned}
&\inf_{P\in\mathcal{P}_{\mathbf{X},\mathbf{Y},\mathbf{Z}}} \mathbf{E}_{P_{\mathbf{X},\mathbf{Y}}}\mathbf{E}_{P_\mathbf{Z}}\Big[c_X\Big(\mathbf{X},F\big(G_a(\mathbf{Z_a}),G_y(\mathbf{Z_y})\big)\Big)+c_Y\Big(\mathbf{Y},D\big(G_y(\mathbf{Z_y})\big)\Big)\Big] \\
=&\inf_{Q_\mathbf{Z}=P_\mathbf{Z}} \mathbf{E}_{P_{\mathbf{X},\mathbf{Y}}}\mathbf{E}_{Q(\mathbf{Z}|\mathbf{X})}\Bigg[c_X\Big(\mathbf{X},F\big(G_a(\mathbf{Z_a}),G_y(\mathbf{Z_y})\big)\Big)+c_Y\Big(\mathbf{Y},D\big(G_y(\mathbf{Z_y})\big)\Big)\Bigg].
\end{aligned}
\tag{9}
$$

■

A.2 From Unimodal to Multimodal

The proof is similar to Proposition 2, and we present a sketch to it. We can first show $P(\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}}|\mathbf{Z}=z)$ are Dirac for all $z\in\mathcal{Z}$. Then we use the fact that $c\Big((\mathbf{X}_{1:M},\mathbf{Y}),(\hat{\mathbf{X}}_{1:M},\hat{\mathbf{Y}})\Big) = \sum_{i=1}^{M} c_{Xi}\Big(\mathbf{X}_i,\hat{\mathbf{X}}_i\Big)+c_Y\Big(\mathbf{Y},\hat{\mathbf{Y}}\Big)$. Finally, we follow the tower rule of expectation and the conditional independence property similar to the proof in Proposition 2 and this concludes the proof.

■

B Encoder and Decoder Design

Fig. 5 illustrates how MFM operates on multimodal time series data. The encoder $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$ can be parametrized by any model that performs multimodal fusion [40, 68]. We choose the Memory Fusion Network (MFN) [68] as our encoder $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$. We use encoder LSTM networks and decoder LSTM networks [10] to parametrize functions $Q(\mathbf{Z_a}_{1:M}|\mathbf{X}_{1:M})$ and $F_{1:M}$ respectively, and FCNNs to parametrize functions G_y, $G_{a\{1:M\}}$ and D.

C Multimodal Features

For each of the multimodal time series datasets as mentioned in Section 3.3, we extracted the following multimodal features: **Language:** We use pre-trained word embeddings (glove.840B.300d) [42] to convert the video transcripts into a sequence of 300 dimensional word vectors. **Visual:** We use Facet [24] to extract a set of features including per-frame basic and advanced emotions and facial action units as indicators of facial muscle movement [14, 13]. **Acoustic:** We use COVAREP [12] to extract low level acoustic features including 12 Mel-frequency cepstral coefficients (MFCCs), pitch tracking and voiced/unvoiced segmenting features, glottal source parameters, peak slope parameters and maxima dispersion quotients. To reach the same time alignment between different modalities we choose the granularity of the input to be at the level of words. The words are aligned with audio using P2FA [66] to get their exact utterance times. We use expected feature values across the entire word for visual and acoustic features since they are extracted at a higher frequencies.

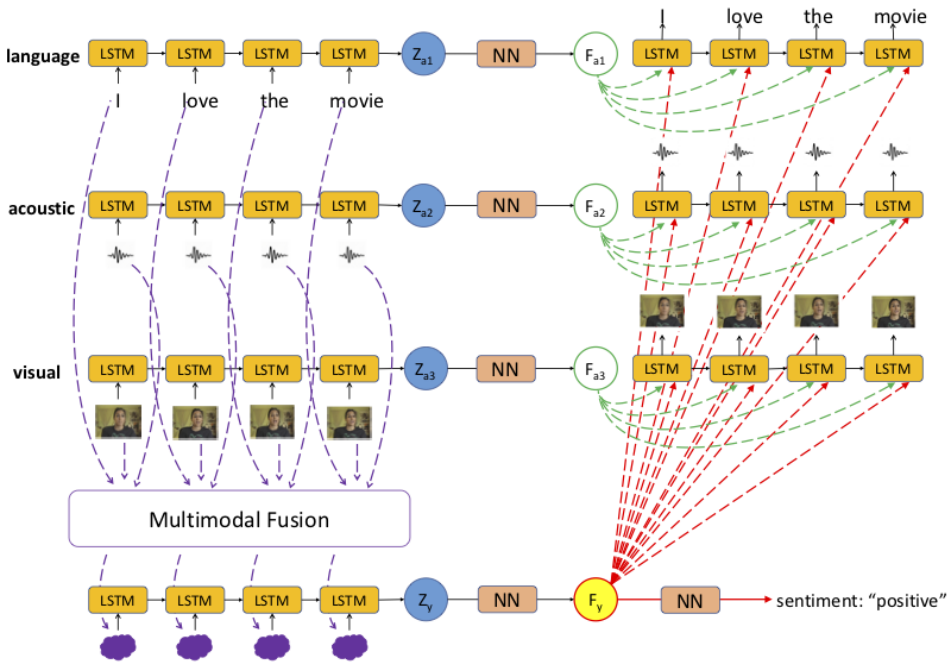


Figure 5: Recurrent neural architecture for MFM. The encoder $Q(\mathbf{Z_y}|\mathbf{X}_{1:M})$ can be parametrized by any model that performs multimodal fusion [40, 68]. We use encoder LSTM networks and decoder LSTM networks [10] to parametrize functions $Q(\mathbf{Z_{a}}_{1:M}|\mathbf{X}_{1:M})$ and $F_{1:M}$ respectively, and FCNNs to parametrize functions G_y, $G_{a\{1:M\}}$ and D.

We make a note that the features for some of these datasets are constantly being updated. The authors of [68] notified us of a discrepancy in the sampling rate for acoustic feature extraction in the ICT-MMMO, YouTube and MOUD datasets which led to inaccurate word-level alignment between the three modalities. They publicly released the updated multimodal features. We performed all experiments on the latest versions of these datasets which can be accessed from `https://github.com/A2Zadeh/CMU-MultimodalSDK`. All baseline models were retrained with extensive hyperparameter search for fair comparison.

D Baseline Models

For a detailed description of the baselines, we point the reader to MFN [68], MARN [69], TFN [67], BC-LSTM [47], MV-LSTM [49], EF-LSTM [22, 17, 52], DF [40], MV-HCRF [55, 56], EF-HCRF [48, 37], THMM [36], SVM-MD [70] and RF [3].

We use the following extra notations for full descriptions of the baseline models described in Subsection 3.3:

Variants of EF-LSTM: **EF-LSTM** (Early Fusion LSTM) uses a single LSTM [22] on concatenated multimodal inputs. We also implement the **EF-SLSTM** (stacked) [17], **EF-BLSTM** (bidirectional) [52] and **EF-SBLSTM** (stacked bidirectional) versions.

Variants of EF-HCRF: **EF-HCRF**: (Hidden Conditional Random Field) [48] uses a HCRF to learn a set of latent variables conditioned on the concatenated input at each time step. **EF-LDHCRF** (Latent Discriminative HCRFs) [37] are a class of models that learn hidden states in a CRF using a latent code between observed concatenated input and hidden output. **EF-HSSHCRF**: (Hierarchical Sequence Summarization HCRF) [56] is a layered model that uses HCRFs with latent variables to learn hidden spatio-temporal dynamics.

Variants of MV-HCRF: **MV-HCRF**: Multi-view HCRF [55] is an extension of the HCRF for Multi-view data, explicitly capturing view-shared and view specific sub-structures. **MV-LDHCRF**: [37] is a

variation of the MV-HCRF model that uses LDHCRF instead of HCRF. **MV-HSSHCRF**: [56] further extends **EF-HSSHCRF** by performing Multi-view hierarchical sequence summary representation.

E Full Results

Prediction: In the following, we provide the full results for all baselines models described in Subsection 3.3. Table 3 contains results for multimodal speaker traits recognition on the POM dataset. Table 4 contains results for the multimodal sentiment analysis on the CMU-MOSI, ICT-MMMO, YouTube, and MOUD datasets. Table 5 contains results for multimodal emotion recognition on the IEMOCAP dataset. MFM consistently achieves state-of-the-art or competitive results for all six multimodal datasets. We believe that by our MFM design, the multimodal discriminative factor $\mathbf{F_y}$ has successfully learned more meaningful representations by distilling discriminative features. This highlights the benefit of learning factorized multimodal representations towards discriminative tasks.

Table 3: Results for personality trait recognition on the POM dataset. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art. Improvements are highlighted in green. MFM achieves state-of-the-art or competitive performance on all datasets and metrics.

Dataset	POM Speaker Personality Traits															
Task	Con	Pas	Voi	Dom	Cre	Viv	Exp	Ent	Res	Tru	Rel	Out	Tho	Ner	Per	Hum
Metric	r															
Majority	-0.041	-0.029	-0.104	-0.031	-0.122	-0.044	-0.065	-0.105	0.006	-0.077	-0.024	-0.085	-0.130	0.097	-0.127	-0.069
SVM	0.063	0.086	-0.004	0.141	0.113	0.076	0.134	0.141	0.166	0.168	0.104	0.066	0.134	0.068	0.064	0.147
DF	0.240	0.273	0.017	0.139	0.112	0.173	0.118	0.217	0.148	0.143	0.019	0.093	0.041	0.136	0.168	0.259
EF-LSTM	0.200	0.302	0.031	0.079	0.170	0.244	0.265	0.240	0.142	0.062	0.083	0.152	0.260	0.105	0.217	0.227
EF-SLSTM	0.221	0.327	0.042	0.151	0.177	0.239	0.268	0.248	0.204	0.069	0.092	0.215	0.252	0.159	0.218	0.196
EF-BLSTM	0.162	0.289	-0.034	0.135	0.191	0.279	0.274	0.231	0.184	0.154	0.093	0.147	0.245	0.166	0.243	0.272
EF-SBLSTM	0.174	0.310	0.021	0.088	0.170	0.224	0.261	0.241	0.155	0.163	0.097	0.120	0.215	0.121	0.216	0.171
MV-LSTM	0.358	0.416	0.131	0.146	0.280	0.347	0.323	0.326	0.295	0.237	0.119	0.238	0.284	0.258	0.239	0.317
BC-LSTM	0.359	0.425	0.081	0.234	0.358	0.417	0.450	0.361	0.293	0.109	0.075	0.078	0.363	0.184	0.344	0.319
TFN	0.089	0.201	0.030	0.020	0.124	0.204	0.171	0.223	-0.051	-0.064	0.114	0.060	0.048	-0.002	0.106	0.213
MARN	0.340	0.410	0.166	0.235	0.340	0.374	0.406	0.378	0.282	0.147	0.215	0.204	0.348	0.235	0.303	0.287
MFN	0.395	0.428	0.193	0.313	0.367	0.431	0.452	0.395	0.333	0.296	0.255	0.259	0.381	0.318	0.377	0.386
MFM	**0.431**	**0.450**	**0.197**	**0.411**	**0.380**	**0.448**	**0.467**	**0.452**	**0.368**	0.212	**0.309**	**0.333**	**0.404**	**0.333**	0.334	**0.408**
Δ_{SOTA}	↑ 0.036	↑ 0.022	↑ 0.004	↑ 0.097	↑ 0.013	↑ 0.017	↑ 0.015	↑ 0.057	↑ 0.035	–	↑ 0.054	↑ 0.074	↑ 0.023	↑ 0.015	–	↑ 0.022

F Information and Gradient-Based Interpretation

Information-Based Interpretation: We choose the normalized Hilbert-Schmidt Independence Criterion [19, 64] as the approximation (see [58, 64]) of our MI measure:

$$\text{MI}(\mathbf{F}_{\cdot}, \hat{\mathbf{X}}_i) = \text{HSIC}_{norm}(\mathbf{F}_{\cdot}, \hat{\mathbf{X}}_i) = \frac{\text{tr}(\mathbf{K}_{\mathbf{F}_{\cdot}}\mathbf{H}\mathbf{K}_{\hat{\mathbf{X}}_i}\mathbf{H})}{\|\mathbf{H}\mathbf{K}_{\mathbf{F}_{\cdot}}\mathbf{H}\|_F \|\mathbf{H}\mathbf{K}_{\hat{\mathbf{X}}_i}\mathbf{H}\|_F}, \tag{10}$$

where $\cdot$ represents y or a_i, n is the number of $\{\mathbf{F}_{\cdot}, \hat{\mathbf{X}}_i\}$ pairs, $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, $\mathbf{K}_{\mathbf{F}_{\cdot}} \in \mathbb{R}^{n\times n}$ is the Gram matrix of $\mathbf{F}_{\cdot}$ with $\mathbf{K}_{\mathbf{F}_{\cdot}ij} = k_1(\mathbf{F}_{\cdot i}, \mathbf{F}_{\cdot j})$, $\mathbf{K}_{\hat{\mathbf{X}}_i} \in \mathbb{R}^{n\times n}$ is the Gram matrix of $\hat{\mathbf{X}}_i$ with $\mathbf{K}_{\hat{\mathbf{X}}_i jk} = k_2(\hat{\mathbf{X}}_{ij}, \hat{\mathbf{X}}_{ik})$. $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are predefined kernel functions.

The most common choice for the kernel is the RBF kernel. However, if we consider time series data with various time steps, we need to either perform data augmentation or choose another kernel choice. For example, we can adopt the Global Alignment Kernel [11] which considers the alignment between two varying-length time series when computing the kernel score between them. To simplify our analysis, we choose to augment data before we calculate the kernel score with the RBF kernel. More specifically, we perform averaging over time series data:

$$\mathbf{X}_{aug} = \frac{1}{n}\sum_{t=1}^{T} X^t \text{ with } \mathbf{X} = [X^1, X^2, \cdots, X^T]. \tag{11}$$

The bandwidth of the RBF kernel is set as 1.0 throughout the experiments.

Table 4: Sentiment prediction results on CMU-MOSI, ICT-MMMO, YouTube and MOUD. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous state of the art (SOTA). Improvements are highlighted in green. MFM achieves state-of-the-art or competitive performance on all datasets and metrics.

Dataset	**CMU-MOSI**					**ICT-MMMO**		**YouTube**		**MOUD**	
Task	Sentiment					Sentiment		Sentiment		Sentiment	
Metric	Acc_7	Acc_2	F1	MAE	r	Acc_2	F1	Acc_3	F1	Acc_2	F1
Majority	17.5	50.2	50.1	1.864	0.057	40.0	22.9	42.4	25.2	60.4	45.5
RF	21.3	56.4	56.3	-	-	70.0	69.8	33.3	32.3	64.2	63.3
SVM-MD	26.5	71.6	72.3	1.100	0.559	68.8	68.7	42.4	37.9	59.4	45.5
THMM	17.8	53.8	53.0	-		50.7	45.4	42.4	27.9	61.3	57.0
SAL-CNN	-	73.0	-	-	-	-	-	-	-	-	-
C-MKL	30.2	72.3	72.0	-	-	-	-	-	-	-	-
EF-HCRF	24.6	65.3	65.4	-	-	50.0	50.3	44.1	43.8	54.7	54.7
EF-LDHCRF	24.6	64.0	64.0	-	-	73.8	73.1	45.8	45.0	52.8	49.3
MV-HCRF	22.6	44.8	27.7	-	-	36.3	19.3	27.1	19.7	60.4	45.5
MV-LDHCRF	24.6	64.0	64.0	-	-	68.8	67.1	44.1	44.0	53.8	46.9
CMV-HCRF	22.3	44.8	27.7	-	-	36.3	19.3	30.5	14.3	60.4	45.5
CMV-LDHCRF	24.6	63.6	63.6	-	-	51.3	51.4	42.4	42.0	53.8	47.8
EF-HSSHCRF	24.6	63.3	63.4	-	-	50.0	51.3	37.3	35.6	52.8	49.3
MV-HSSHCRF	24.6	65.6	65.7	-	-	62.5	63.1	44.1	44.0	47.2	46.4
DF	26.8	72.3	72.1	1.143	0.518	65.0	58.7	45.8	32.0	67.0	67.1
EF-LSTM	32.4	74.3	74.3	1.023	0.622	66.3	65.0	44.1	43.6	67.0	64.3
EF-SLSTM	29.3	72.7	72.8	1.081	0.600	72.5	70.9	40.7	41.2	56.6	51.4
EF-BLSTM	28.9	72.0	72.0	1.080	0.577	63.8	49.6	42.4	38.1	58.5	58.9
EF-SBLSTM	26.8	73.3	73.2	1.037	0.619	62.5	49.0	37.3	33.2	63.2	63.3
MV-LSTM	33.2	73.9	74.0	1.019	0.601	72.5	72.3	45.8	43.3	57.6	48.2
BC-LSTM	28.7	73.9	73.9	1.079	0.581	70.0	70.1	45.0	45.1	72.6	72.9
TFN	28.7	74.6	74.5	1.040	0.587	72.5	72.6	45.0	41.0	63.2	61.7
MARN	34.7	77.1	77.0	0.968	0.625	71.3	70.2	48.3	44.9	81.1	81.2
MFN	34.1	77.4	77.3	0.965	0.632	73.8	73.1	51.7	51.6	81.1	80.4
MFM	**36.2**	**78.1**	**78.1**	**0.951**	**0.662**	**81.3**	**79.2**	**53.3**	**52.4**	**82.1**	**81.7**
Δ_{SOTA}	↑ 1.5	↑ 0.7	↑ 0.8	↓ 0.014	↑ 0.030	↑ 7.5	↑ 6.1	↑ 1.6	↑ 0.8	↑ 1.0	↑ 0.5

Here, we provide an additional interpretation result for the POM dataset in Table 6. We observe that the language modality is also the most informative while the visual and acoustic modalities are almost equally informative. This result is in agreement with behavioral studies which have observed that non-verbal behaviors are particularly informative of personality traits [20, 31, 35]. For example, the same sentence "this movie was great" can convey significantly different messages on speaker confidence depending on whether it was said in a loud and exciting voice, with eye contact, or powerful gesticulation.

Gradient-Based Interpretation: MFM reconstructs x_i as follows:

$$\hat{x}_i = F_i(f_{ai}, f_y), f_{ai} = G_{ai}(z_{ai}), f_y = G_y(z_y), z_{ai} \sim Q(\mathbf{Z_{ai}}|\mathbf{X}_i = x_i), z_y \sim Q(\mathbf{Z_y}|\mathbf{X}_{1:M} = x_{1:M}). \quad (12)$$

Equation equation 12 also explains how we obtain $f_y \sim P(\mathbf{F_y}|\mathbf{X}_{1:M} = x_{1:M})$. The gradient flow through time is defined as:

$$\nabla_{f_y}(\hat{x}_i) := [\|\nabla_{f_y}\hat{x}_i^1\|_F^2, \|\nabla_{f_y}\hat{x}_i^2\|_F^2, \cdots, \|\nabla_{f_y}\hat{x}_i^T\|_F^2]. \quad (13)$$

G Surrogate Inference Graphical Model

We illustrate the surrogate inference for addressing the missing modalities issue in Figure 6. The surrogate inference model infers the latent codes given the present modalities. These inferred latent codes can then be used for reconstructing the missing modalities or label prediction in the presence of missing modalities.

H Comparison with [23]

A similar approach for factorizing the latent factors was recently proposed by [23] in work that was performed independently and in parallel. In comparison with MFM, there are several major

Table 5: Emotion recognition results on IEMOCAP test set. The best results are highlighted in bold and Δ_{SOTA} shows the change in performance over previous SOTA. Improvements are highlighted in green. MFM achieves state-of-the-art or competitive performance on all datasets and metrics.

Dataset	**IEMOCAP Emotions**											
Task	Happy		Sad		Angry		Frustrated		Excited		Neutral	
Metric	Acc_2	F1	Acc_2	F1	Acc_2	F1	Acc_2	F1	Acc_2	F1	Acc_2	F1
Majority	85.6	79.0	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
SVM	86.1	81.5	81.1	78.8	82.5	82.4	77.3	71.1	86.4	86.0	65.2	64.9
RF	85.5	80.7	80.1	76.5	81.9	82.0	78.6	75.3	88.9	85.1	63.2	57.3
THMM	85.6	79.2	79.5	79.8	79.3	73.0	71.6	69.6	86.0	84.6	58.6	46.4
EF-HCRF	85.7	79.2	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
EF-LDHCRF	85.8	79.5	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
MV-HCRF	15.0	4.9	79.4	70.3	24.2	9.4	79.5	70.4	89.6	84.7	59.1	44.0
MV-LDHCRF	85.7	79.2	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
CMV-HCRF	14.4	3.6	79.4	70.3	24.2	9.4	79.5	70.4	89.6	84.7	59.1	44.0
CMV-LDHCRF	85.8	79.5	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
EF-HSSHCRF	85.8	79.5	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
MV-HSSHCRF	85.8	79.5	79.4	70.3	75.8	65.4	79.5	70.4	89.6	84.7	59.1	44.0
DF	86.0	81.0	81.8	81.2	75.8	65.4	78.4	76.8	89.6	84.7	59.1	44.0
EF-LSTM	85.2	83.3	82.1	81.1	84.5	84.3	79.5	70.4	89.6	84.7	68.2	67.1
EF-SLSTM	85.6	79.0	80.7	80.2	82.8	82.2	77.5	69.7	89.3	86.2	68.8	**68.5**
EF-BLSTM	85.0	83.7	81.8	81.6	84.2	83.3	79.5	70.4	89.6	84.7	67.1	66.6
EF-SBLSTM	86.0	84.2	80.2	80.5	85.2	84.5	79.5	70.4	89.6	84.7	67.8	67.1
MV-LSTM	85.9	81.3	80.4	74.0	85.1	84.3	79.5	73.8	88.9	85.8	67.0	66.7
BC-LSTM	84.9	81.7	83.2	81.7	83.5	84.2	80.0	76.1	86.9	85.4	67.5	64.1
TFN	84.8	83.6	83.4	82.8	83.4	84.2	74.1	74.3	75.6	78.0	67.5	65.4
MARN	86.7	83.6	82.0	81.2	84.6	84.2	79.5	76.6	89.6	**87.1**	66.8	65.9
MFN	90.1	85.3	85.8	79.2	87.0	86.0	80.3	**76.9**	89.8	86.3	71.8	61.7
MFM	**90.2**	**85.8**	**88.4**	**86.1**	**87.5**	**86.7**	**80.4**	74.5	**90.0**	**87.1**	**72.1**	68.1
Δ_{SOTA}	↑ 0.1	↑ 0.5	↑ 2.6	↑ 3.3	↑ 0.5	↑ 0.7	↑ 0.1	–	↑ 0.2	–	↑ 0.3	–

Table 6: Information-Based interpretation results showing ratios $r_i = \frac{\text{MI}(\mathbf{F_y}, \hat{\mathbf{X}}_i)}{\text{MI}(\mathbf{F}_{\mathbf{a}i}, \hat{\mathbf{X}}_i)}$, $i \in \{(\ell)anguage, (v)isual, (a)coustic\}$ for the POM dataset for personality traits prediction.

Ratio	r_ℓ (language)	r_v (visual)	r_a (acoustic)
POM	1.090	0.996	0.898

differences that can be categorized into the (1) prior matching discrepancy, (2) inference network, (3) discriminative objective, (4) multimodal fusion, (5) scale of experiments.

1. MFM uses $\mathcal{MMD}(Q_{\mathbf{Z}}, P_{\mathbf{Z}})$ (see Equation 4) as the prior matching discrepancy while [23] use $\mathcal{KL}(Q_{\mathbf{Z}|\mathbf{X}}, P_{\mathbf{Z}})$) (see Section 2.2.1 in [23]).
2. MFM considers multimodal and unimodal inference in a single network (see Figure 1(b)), while [23] considers separate networks (see Figure 1 in [23]). They further propose to match the coherence between these two networks using an additional loss term (see Equation 7 in [23]).
3. MFM learns to predict the labels using a generative framework (see Figure 1 (a)), while [23] use an additional hinge loss to separate the latent factors from different labels (see Equation (9) in [23]).
4. MFM is a flexible framework that can be combined with any multimodal fusion encoder (see Section 3.1), while [23] considers a fixed multimodal encoder (similar to early fusion) (see Section 4.1 in [23]).
5. We evaluate the performance of MFM over a much larger scale of datasets. We perform experiments on six multimodal time-series datasets that take on the form of videos with the language, visual, and acoustic modalities. These datasets span three core research areas of multimodal personality traits recognition, multimodal sentiment analysis, and multimodal emotion recognition. On the other hand, [23] evaluates their model on a spoken digit dataset

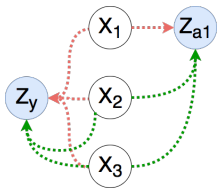


Figure 6: The surrogate inference graphical model to deal with missing modalities in MFM. Red lines denote original inference in MFM and green lines denote surrogate inference to infer latent codes given present modalities.

which randomly combines a digit image with a spoken digit (see Section 4 in [23]). MFM further considers experiments to evaluate reconstruction and prediction in the presence of missing modalities (see Section 3.3) which [23] do not. Lastly, we compares to over 20 baseline models in our experiments (see Section 3.3) and explore the choice of various multimodal encoders in MFM. [23] only compares to the JMVAE baseline model which resembles the $\mathbf{M_D}$ model in our ablation study (see Section 4.4 in [23]).

In Table 7, we provide a comparison on CMU-MOSI to test the disentanglement performance for the model described in [23]. MFM with early fusion encoder and MFN encoder both perform better than the model proposed in [23].

Table 7: Comparison with [23] for CMU-MOSI sentiment analysis.

Metric (encoder design)	Acc_7	Acc_2	F1	MAE	r
without factorization (early fusion)	32.4	74.3	74.3	1.023	0.622
without factorization (MFN)	34.1	77.4	77.3	0.965	0.632
[23]	33.8	75.2	75.2	1.049	0.584
MFM (early fusion)	32.8	75.2	75.3	1.020	0.629
MFM (MFN)	**36.2**	**78.1**	**78.1**	**0.951**	**0.662**