
Appendix for FCD

Anonymous Author(s)

Affiliation

Address

email

A Proof of Theorem 4.1

We first give the following lemma, which describes the probability distribution between two variables X and Y .

Lemma A.1. *For two variables X, Y defined on the measurable sets \mathcal{X}, \mathcal{Y} with the same measure, and the probability distributions are $P(X), P(Y)$. If $Y = \mathcal{F}(X)$, where \mathcal{F} is a measure-preserving bijective function, we have $P(X) = P(Y)$.*

Proof. Since \mathcal{F} is a measure-preserving bijective function, we have:

$$P(X \in \mathcal{X}) = P(\mathcal{F}^{-1}(Y) \in \mathcal{X}) = P(Y \in \mathcal{F}(\mathcal{X})) = P(Y \in \mathcal{Y}) \quad (1)$$

□

Corollary A.2. *For two variables X, Y from measurable sets \mathcal{X}, \mathcal{Y} with their measures $\mu(\mathcal{X})$ and $\mu(\mathcal{Y})$. If $Y = \mathcal{F}(X)$ and \mathcal{F} is a measure-preserving bijective function, there exists:*

$$P(X, Y) = P(X)\delta(Y - \mathcal{F}(X)) = P(Y)\delta(Y - \mathcal{F}(X)) \quad (2)$$

when $\mu(\mathcal{X}) = \mu(\mathcal{Y})$.

where $\delta(t)$ is the standard Dirac delta function which has the following properties for $t \in \mathbb{R}$:

$$1. \delta(t) = 0 \text{ for all } t \neq 0$$

$$2. \int_{-\infty}^{+\infty} \delta(t)dt = 1.$$

Proof. Based on Bayes' theorem, we have:

$$P(X, Y) = P(X)P(Y|X) \quad (3)$$

Since \mathcal{F} is a bijective function, Y is uniquely determined by X , according to [6], $P(Y|X) = \delta(Y - \mathcal{F}(X))$. Based on Lemma A.1, we have:

$$P(X, Y) = P(X)P(Y|X) = P(X)\delta(Y - \mathcal{F}(X)) = P(Y)\delta(Y - \mathcal{F}(X)) \quad (4)$$

□

Corollary A.3. *For three variables X, Y and Z defined on the measurable sets \mathcal{X}, \mathcal{Y} and \mathcal{Z} , and the joint probability distributions are $P(X, Z), P(Y, Z)$. If $Y = \mathcal{F}(X)$, where \mathcal{F} is a measure-preserving bijective function, X is independent with Z , and Y is also independent with Z , we have $P(X, Z) = P(Y, Z)$.*

Proof. Let $\mu(\mathcal{X})$ is the measure of \mathcal{X} . Since $Y = \mathcal{F}(X)$, where \mathcal{F} is a measure-preserving bijective function, then \mathcal{X} and \mathcal{Y} have the same measure, i.e., $\mu(\mathcal{X}) = \mu(\mathcal{Y})$. Besides, the measures of the

joint domains $\mathcal{X} \times \mathcal{Z}$, $\mathcal{Y} \times \mathcal{Z}$ can be formed as $\mu(\mathcal{X} \times \mathcal{Z})$ and $\mu(\mathcal{Y} \times \mathcal{Z})$, respectively. Based on the Fubini's Theorem, we have:

$$\mu(\mathcal{X} \times \mathcal{Z}) = \mu(\mathcal{X})\mu(\mathcal{Z}) = \mu(\mathcal{Y})\mu(\mathcal{Z}) = \mu(\mathcal{Y} \times \mathcal{Z}) \quad (5)$$

Thus, the measures of joint domains $\mathcal{X} \times \mathcal{Z}$ and $\mathcal{Y} \times \mathcal{Z}$ are the same. Since \mathcal{F} is a measure-preserving bijective function, X is independent with Z , and Y is also independent with Z , based on Lemma A.1, we have:

$$\begin{aligned} P((X, Z) \in (\mathcal{X} \times \mathcal{Z})) &= P(X \in \mathcal{X})P(Z) \\ &= P(\mathcal{F}^{-1}(Y) \in \mathcal{X})P(Z) = P(Y \in \mathcal{F}(\mathcal{X}))P(Z) = P((Y, Z) \in (\mathcal{Y} \times \mathcal{Z})) \end{aligned} \quad (6)$$

□

Corollary A.4. For three variables X, Y, Z from measurable sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, where X is independent with Z and Y is also independent with Z . If $Y = \mathcal{F}(X)$ and \mathcal{F} is a measure-preserving bijective function, there exists:

$$P(X, Y, Z) = P(X, Z)\delta(Y - \mathcal{F}(X)) = P(Y, Z)\delta(Y - \mathcal{F}(X)) \quad (7)$$

when $\mu(\mathcal{X}) = \mu(\mathcal{Y})$.

Proof. Based on Bayes' theorem, we have:

$$P(X, Y, Z) = P(X, Z)P(Y|X, Z) \quad (8)$$

Since \mathcal{F} is a bijective function, Y is uniquely determined by X . Thus, X is independent with Z and Y is also independent with Z , then we have:

$$P(Y|X, Z) = \frac{P(Y, X, Z)}{P(X, Z)} = \frac{P(Y, X)P(Z)}{P(X)P(Z)} = \frac{P(Y, X)}{P(X)} = P(Y|X)$$

According to Corollary A.2 and Corollary A.3, and $P(Y|X)$ is a Dirac delta function $\delta(Y - \mathcal{F}(X))$, we have:

$$\begin{aligned} P(X, Y, Z) &= P(X, Z)P(Y|X, Z) \\ &= P(X, Z)P(Y|X) = P(X, Z)\delta(Y - \mathcal{F}(X)) = P(Y, Z)\delta(Y - \mathcal{F}(X)) \end{aligned} \quad (9)$$

□

Based on Corollary A.2 and Corollary A.4, we can prove the Theorem 4.1 as follows.

Proof. For the hidden features \mathbf{h}_n^m , \mathbf{u}_n^m and the ground truth y_n , we are seeking to maximizing the expectation of the conditional mutual information between \mathbf{u}_n^m and the ground truth y_n , when \mathbf{h}_n^m is given. From the probability perspective, the \mathbf{h}_n^m and \mathbf{u}_n^m are both independent with the given ground truth y_n . Based on the definition of conditional mutual information and the Bayes' theorem, we have:

$$\begin{aligned} \max \mathbb{E}_{P(\mathbf{h}_n^m)} [I(\mathbf{u}_n^m; y_n | \mathbf{h}_n^m)] &= \max \mathbb{E}_{P(\mathbf{h}_n^m)} \left[\mathbb{E}_{P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m)} \left[\log \frac{P(\mathbf{u}_n^m, y_n | \mathbf{h}_n^m)}{P(\mathbf{u}_n^m | \mathbf{h}_n^m)P(y_n | \mathbf{h}_n^m)} \right] \right] \\ &= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(\mathbf{h}_n^m) P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n | \mathbf{h}_n^m)}{P(\mathbf{u}_n^m | \mathbf{h}_n^m)P(y_n | \mathbf{h}_n^m)} dy_n d\mathbf{h}_n^m d\mathbf{u}_n^m d\mathbf{h}_n^m \\ &= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(\mathbf{h}_n^m) P(y_n | \mathbf{u}_n^m, \mathbf{h}_n^m) P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n | \mathbf{h}_n^m)}{P(\mathbf{u}_n^m | \mathbf{h}_n^m)P(y_n | \mathbf{h}_n^m)} dy_n d\mathbf{h}_n^m d\mathbf{u}_n^m d\mathbf{h}_n^m \\ &= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(\mathbf{h}_n^m) P(y_n | \mathbf{u}_n^m, \mathbf{h}_n^m) P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n | \mathbf{h}_n^m) P(\mathbf{h}_n^m)}{P(\mathbf{u}_n^m | \mathbf{h}_n^m) P(y_n | \mathbf{h}_n^m) P(\mathbf{h}_n^m)} dy_n d\mathbf{h}_n^m d\mathbf{u}_n^m d\mathbf{h}_n^m \\ &= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(\mathbf{h}_n^m) P(y_n | \mathbf{u}_n^m, \mathbf{h}_n^m) P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n, \mathbf{h}_n^m)}{P(\mathbf{u}_n^m, \mathbf{h}_n^m) P(y_n | \mathbf{h}_n^m)} dy_n d\mathbf{h}_n^m d\mathbf{u}_n^m d\mathbf{h}_n^m \end{aligned} \quad (10)$$

Note that the distribution of the ground truth $P(y_n)$ doesn't change with \mathbf{h}_n^m , thus we have $P(y_n | \mathbf{h}_n^m) = P(y_n)$. Besides, based on the core idea of backdoor-adjustment that causal intervention is applied for the confounder to cut off the causal relation between the treatment and the confounder [9], we apply the causal intervention on \mathbf{h}_n^m to turn it into its counterfactual form $\mathbf{h}_n^{m'}$,

and cut off the causal relation between \mathbf{h}_n^m and \mathbf{u}_n^m . From Corollary A.2 and Corollary A.4, the Equation (10) can be further transformed to Equation (11).

$$\begin{aligned}
&= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(\mathbf{h}_n^m) P(y_n | \mathbf{u}_n^m, \mathbf{h}_n^m) P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n) \delta(\mathbf{u}_n^m - \mathcal{F}_{\text{mb}}^m(\mathbf{h}_n^m; \theta_{\text{mb}}^m))}{P(\mathbf{u}_n^m) \delta(\mathbf{u}_n^m - \mathcal{F}_{\text{mb}}^m(\mathbf{h}_n^m; \theta_{\text{mb}}^m)) P(y_n)} dy_n d\mathbf{h}_n^m d\mathbf{u}_n^m d\mathbf{h}_n^m \\
&\xrightarrow{\text{Causal Intervention}} \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}} \underbrace{P(\mathbf{h}_n^{m'}) P(y_n | \mathbf{u}_n^m, \mathbf{h}_n^{m'})}_{\text{Backdoor-adjustment}} P(\mathbf{u}_n^m, \mathbf{h}_n^m) \log \frac{P(\mathbf{u}_n^m, y_n)}{P(\mathbf{u}_n^m) P(y_n)} dy_n d\mathbf{h}_n^m ddo(\mathbf{u}_n^m) d\mathbf{h}_n^{m'} \\
&= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \underbrace{P(\mathbf{h}_n^{m'}) P(y_n | \mathbf{u}_n^m, \mathbf{h}_n^{m'})}_{\text{Backdoor-adjustment}} d\mathbf{h}_n^{m'} \int_{\mathbb{R}^d} P(\mathbf{u}_n^m, \mathbf{h}_n^m) d\mathbf{h}_n^m \log \frac{P(\mathbf{u}_n^m, y_n)}{P(\mathbf{u}_n^m) P(y_n)} dy_n d\mathbf{u}_n^m \\
&= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(y_n | do(\mathbf{u}_n^m)) P(do(\mathbf{u}_n^m)) \log \frac{P(do(\mathbf{u}_n^m), y_n)}{P(do(\mathbf{u}_n^m)) P(y_n)} dy_n ddo(\mathbf{u}_n^m) \\
&= \max \int_{\mathbb{R}^d} \int_{\mathbb{R}} P(do(\mathbf{u}_n^m), y_n) \log \frac{P(do(\mathbf{u}_n^m), y_n)}{P(do(\mathbf{u}_n^m)) P(y_n)} dy_n ddo(\mathbf{u}_n^m) \\
&= \max I(do(\mathbf{u}_n^m); y_n).
\end{aligned} \tag{11}$$

where $\mathcal{F}_{\text{mb}}^m(\cdot; \theta_{\text{mb}}^m)$ is the bijective decomposition function with parameter θ_{mb}^m (the URD module in our paper). $do(\cdot)$ is *do*-operator for causal intervention in backdoor-adjustment. $\mathbf{h}_n^{m'}$ is the \mathbf{h}_n^m after causal intervention. From Equation (10) and Equation (11), Theorem 4.1 holds. \square

B Pseudo Code for Training Process

We first give the training process to show the reversibility of $\mathcal{F}_{\text{URD-d}}^m$ and $\mathcal{F}_{\text{URD-f}}^m$.

Algorithm 1: Enforce Full Rank in PyTorch style

Input: Multimodal model with FCD Model

```

1 for  $m \leftarrow 1$  to  $M$  do
    // Get parameters of linear layers in URD
2    $\theta_{\text{URD-d}}^m, \theta_{\text{URD-f}}^m \leftarrow \text{get}(\text{Model.CCD.URD}, \text{Linear});$ 
    // SVD
3    $U_{\text{URD-d}}^m, \Sigma_{\text{URD-d}}^m, V_{\text{URD-d}}^m \leftarrow \text{SVD}(\theta_{\text{URD-d}}^m \cdot \text{weight});$ 
4    $U_{\text{URD-f}}^m, \Sigma_{\text{URD-f}}^m, V_{\text{URD-f}}^m \leftarrow \text{SVD}(\theta_{\text{URD-f}}^m \cdot \text{weight});$ 
    // Enforce Full Rank (Invertible)
5    $\Sigma_{\text{URD-d}}^m \leftarrow \text{torch.maximum}(\Sigma_{\text{URD-d}}^m, 1e^{-5});$ 
6    $\Sigma_{\text{URD-f}}^m \leftarrow \text{torch.maximum}(\Sigma_{\text{URD-f}}^m, 1e^{-5});$ 
    // Reconstruct
7    $\theta_{\text{URD-d}}^m \cdot \text{weight} \leftarrow \text{nn.Parameter}(U_{\text{URD-d}}^m \Sigma_{\text{URD-d}}^m V_{\text{URD-d}}^{m\top});$ 
8    $\theta_{\text{URD-f}}^m \cdot \text{weight} \leftarrow \text{nn.Parameter}(U_{\text{URD-f}}^m \Sigma_{\text{URD-f}}^m V_{\text{URD-f}}^{m\top});$ 
9 end
```

Algorithm 2: Training process of our process method in PyTorch style

Input: Multimodal dataset \mathcal{D} , Multimodal model with FCD Model, Optimizer opt, Number of epochs N_e .

```

1 for  $e$  in  $\{1, \dots, N_e\}$  do
2    $\mathcal{L}_{\text{overall}} \leftarrow \text{Model.forward}(\mathcal{D});$ 
3    $\mathcal{L}_{\text{overall}}.\text{backward}();$ 
4    $\text{opt.step}();$ 
5   Apply Algorithm 1 with model as input;
6 end
```

Besides, to achieve measure-preserving, we apply the following constraint on the weight matrix of $\mathcal{F}_{\text{URD-d}}^m$ and $\mathcal{F}_{\text{URD-f}}^m$. Note that the nesting of two measure-preserving functions is still measure-preserving [12].

Algorithm 3: Measure-preserving Linear Layer Forward

Input: Feature vector \mathbf{x} **Output:** Forward output vector \mathbf{x}'

// Get weight matrix and bias

```
1  $\mathbf{W} \leftarrow \theta_*^m.\text{weight};$   
62 2  $\mathbf{b} \leftarrow \theta_*^m.\text{bias};$   
   // Construct skew symmetric matrix  
3  $\mathbf{A} \leftarrow \mathbf{W} - \mathbf{W}^\top;$   
4  $\mathbf{W}' \leftarrow \text{torch.matrix\_exp}(\mathbf{A});$   
   // Linear forward  
5  $\mathbf{x}' \leftarrow \mathbf{x}\mathbf{W}'^\top + \mathbf{b};$ 
```

63 C Introduction to Datasets and Evaluation Metrics in Our Experiments

64 CMU-MOSI and CMU-MOSEI are popularly used in Multimodal Semantic Analysis task [2; 4] with
65 3 modalities. Each sample from both of these datasets is annotated with a sentiment value ranging
66 from -3 (strongly negative) to $+3$ (strongly positive), indicating the polarity and relative strength of
67 the expressed sentiment within each sample. The former one contains 2199 utterance video segments
68 taken from 93 YouTube, and the latter one contains 22,856 utterance video segments [3]. Table 1
summarizes the training, validation and test subset splits following [5; 17].

Table 1: Datasets splits (train, validation (val), and test) in our experiments.

DATASET	TRAIN	VAL	TEST	OVERALL
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16326	1871	4659	22856
MVSA-SINGLE	1555	518	519	2592
UPMC FOOD101	62971	5000	22715	90686
HFM	19816	2410	2409	24635

69
70 MSVA-Single, UPMC Food101, and HFM datasets only have two modalities. MVSA-Single is a
71 commonly used text-image sentiment dataset collected from Twitter. It has 3 categories: positive,
72 neutral and negative with 1398, 724 and 470 samples, respectively [10]. UPMC Food101 dataset
73 is usually used for multimodal image classification [3], which has 101 categories with about 100k
74 images. HFM has two categories, *i.e.*, positive and negative.

75 We report our results with *the mean absolute error (MAE)*, which is calculated by averaging the
76 absolute value between the predicted and ground truth; *Pearson correlation (Corr)* quantifies the
77 extent to which predictions deviate from a linear relationship; *binary classification accuracy (Acc-2)*
78 and *weighted F1 scores (F1)* are computed for both the negative/non-negative (non-exclude 0)
79 [16] and negative/positive (exclude 0) [11], which is achieved by filtering out the samples whose
80 annotation is 0. Following [3], we report the results on the remaining datasets with *accuracy (Acc)*
81 and *weighted F1 score (F1)*.

82 D Introduction to Quantitative Experiment Methods

83 Due to the page limit, we introduce the methods incorporated in our quantitative experiment here.

- 84 • Self-MM [15] focused on restricted ability in capturing differentiated information within
85 each modality due to the unified multimodal annotation, and proposed a self-supervised
86 multi-task learning method to generate unimodal annotation. Besides, Self-MM shifts the
87 generated annotation according to the relative distance to the class center in each modality
88 space.
- 89 • MMIM [4] designed an MI based method to preserve critical task-related information that
90 flows from the original input to the fusion representation. It employed a tight lower bound
91 of MI and estimated the lower bound via likelihood maximization and Gaussian Mixture

Model (GMM). Then a Contrastive Predictive Coding (CPC) loss was employed to retain the modality-invariant information in fused representation.

- MCL-MCF [2] considered that fusion is a progressive process, and provided a hierarchical structure to maximize the maintenance of semantic information during different fusion level via contrastive learning. Besides, MCL-MCF used 1-D convolutional layers to fuse features at different level.
- AtCAF [5] started from casual inference perspective. It blocked the back-door path between text modality and target annotation via front-door adjustment. Then it applied counterfactual reasoning to the attention matrix integrated in cross-attention fusion process to improve the fusion robustness.
- MMML [14] proposed a Multimodal Multi-Loss Fusion Network that integrates pretrained audio and text encoders, cross- and self-attention mechanisms, and multi-loss training to enhance sentiment analysis. The model achieved state-of-the-art performance on various datasets, demonstrating the effectiveness of multimodal fusion and contextual modeling.
- MMBT [7] proposed a multimodal fusion method for text-image classification based on bitransformer. It jointly finetuned the pretrained unimodal encoders by mapping image embeddings to textual token space.
- CLMLF [8] focused on the token level multimodal fusion and employ contrastive learning to align the representations from different modalities.
- MVCN [13] focused on the challenge of modality heterogeneity in multimodal tasks, and proposed to filter redundant visual features based on sparsemax mechanism. Besides, it calibrated feature shift in representation space by minimizing the intra-class discrepancy.
- URMF [3] adopted a multivariate Gaussian distribution to represent spotty semantic instances in a noisy latent space and tried to eliminate the impact of unimodal aleatoric uncertainty to perform robust multimodal fusion via estimating the Gaussian distribution behind features.

E Hyper-parameter Setting and Sensitive Analysis

The hyper-parameters setting of FCD used in different base methods are reported in Table 2. Since FCD can be integrated into any multi-modal intermediate fusion method, the hyper-parameters may be different with each other. We search the appropriate hyper-parameters with two steps: 1) scale search: ranging each hyper-parameter in $\{0.5, 0.05, 0.005, 0.0005, 0.00005\}$, 2) fine-grained search: ranging each hyper-parameter in the scale that achieves the best performance in step 1). For example, if $\lambda_1 = 0.05$ achieves the best performance in step 1), we then range λ_1 from 0.01 to 0.09 in step 2).

Table 2: Hyper-parameters (λ_1 , λ_2 and λ_3) settings in our experiments.

METHOD	DATASET	λ_1	λ_2	λ_3	METHOD	DATASET	λ_1	λ_2	λ_3
SELF-MM	CMU-MOSI	0.08	0.05	0.005	MMBT	MVSA-S	0.005	0.5	0.05
	CMU-MOSEI	0.003	0.009	0.08		FOOD101	0.5	0.005	0.5
MMIM	CMU-MOSI	0.08	0.07	0.9	CLMLF	MVSA-S	0.02	0.0004	0.02
	CMU-MOSEI	0.6	0.04	0.0003		HFM	0.5	0.5	0.005
MCL-MCF	CMU-MOSI	0.01	0.3	0.09	MVCN	MVSA-S	0.05	0.05	0.005
	CMU-MOSEI	0.007	0.0003	0.8		HFM	0.5	0.0005	0.1
AtCAF	CMU-MOSI	0.09	0.0005	0.3	URMF	MVSA-S	0.5	0.07	0.09
	CMU-MOSEI	0.02	0.0004	0.007		FOOD101	0.05	0.0005	0.05
MMML	CMU-MOSI	0.005	0.05	0.05	-	-	-	-	-
	CMU-MOSEI	0.05	0.0005	0.05		-	-	-	-

Additionally, we report the sensitive analysis of several methods: Self-MM, MMIM, MCL-MCF and AtCAF on CMU-MOSI dataset. We fix other hyper-parameters to the value in Table 2 when a specific hyper-parameter are ranging in the corresponding scale of magnitude.

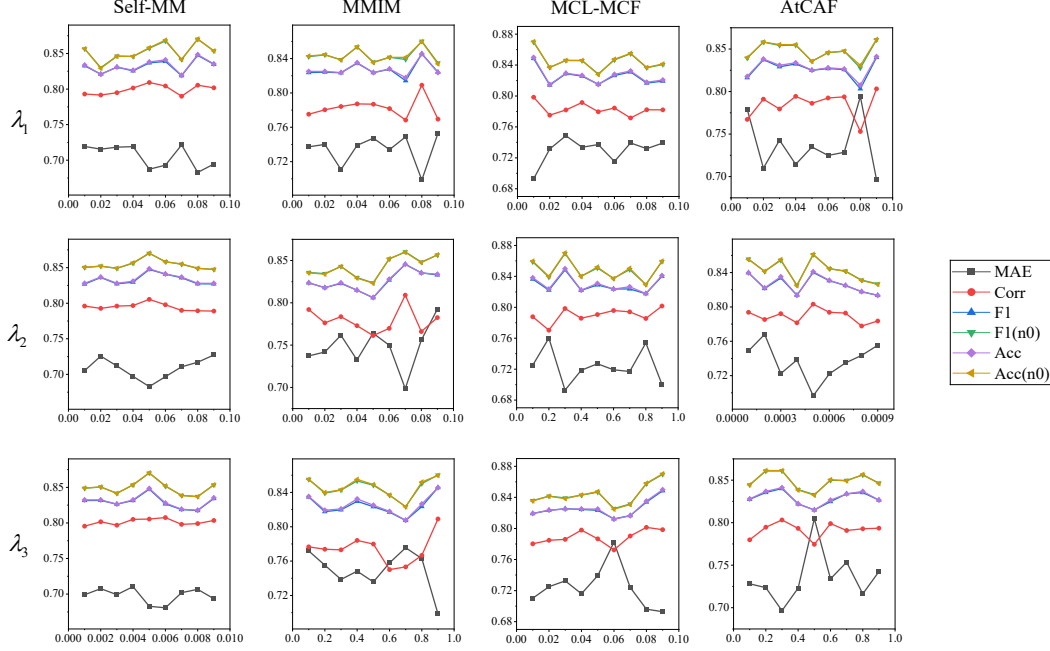


Figure 1: Sensitive analysis of hyper-parameters λ_1 , λ_2 and λ_3 with Self-MM, MMIM, MCL-MCF and AtCAF. The horizontal coordinate axis denotes the hyper-parameter rangings and the vertical coordinate axis denotes the evaluation metrics. "n0" in F1(n0) and Acc(n0) stands for the exclusion of 0 when these two metrics are calculated.

Figure 1 shows the variation tendencies of FCD's prediction performance with the changing of values for the hyper-parameters in Equation (19), *i.e.*, (1) λ_1 : the weight for \mathcal{L}_{MI} which supervises the extraction of unique feature without redundant information, (2) λ_2 : the weight for \mathcal{L}_{dis} which keeps the discriminative information between redundant features and further constrains the modality-specific feature extraction, and (3) λ_3 : the weight for \mathcal{L}_{SDA} which controls the quality of modality-invariant information extraction and synergistic feature alignment. From Figure 1, there seems to be no obvious common trend among different methods for each hyper-parameter. This may be caused by the way how FCD cooperates with each method. Generally, *MAE* shows an opposite trend of change compared to other metrics. For most cases, there exists a significant peak (or valley for *MAE*) of each metric, representing the suitable value. When *MAE* reaches a valley, other metrics reach peaks and vice versa. However, for λ_3 of AtCAF, there exists a peak for *MAE* and valleys for other metrics, which is significantly different with other cases. This may be caused by an inappropriate amplitude of change, where a more fine-grained hyper-parameter search is required. Therefore, FCD needs careful hyper-parameter searching to achieve its greatest potential.

F Computational Overhead Analysis

To further analysis the effectiveness of FCD, we give the following computational overhead analysis to discover training time cost brought by FCD. We conduct this experiment employing Self-MM, MMIM, MCL-MCF, AtCAF, and MMML as the base models to calculate the training overhead per epoch (seconds). The results are shown in Table 3.

In Table 3, we report the average value and the standard deviation of duration to train one epoch on CMU-MOSI and CMU-MOSEI datasets when FCD is applied (Ours (w/ FCD)) or not (Base (w/o FCD)). From Table 3, we can find that the training time of each epoch is different between different methods. This may be caused by the original structure and computation complexity of each base method. When FCD is applied, the increment is also different. This may be caused by the various hidden dimensions, the default batch sizes and the number of modalities (*e.g.*, MMML only has two modalities).

Table 3: The computational time overhead per epoch (seconds) of Self-MM, MMIM, MCL-MCF, AtCAF, and MMML.

METHOD	DATASET	BASE (w/o FCD)	OURS (w/ FCD)
SELF-MM	CMU-MOSI	3.92±0.40	5.12±0.47
	CMU-MOSEI	37.14±3.72	51.13±4.37
MMIM	CMU-MOSI	18.73±1.30	24.98±1.90
	CMU-MOSEI	97.74±4.61	131.57±8.09
MCF-MCL	CMU-MOSI	21.19±1.43	27.45±2.39
	CMU-MOSEI	242.30±7.87	327.76±10.84
AtCAF	CMU-MOSI	17.26±1.69	21.95±1.70
	CMU-MOSEI	200.61±9.15	250.68±9.97
MMML	CMU-MOSI	228.52±1.37	239.12±1.85
	CMU-MOSEI	2558.47±107.736	2570.14±90.25

G Broader Impacts and Future Works

FCD is a plug-and-play module that can be integrated into any existing intermediate multimodal models to handle the unimodal uncertain noise whilst makes full use of the task-related information. We believe that FCD can bring more attentions to current multimodal representation learning community about handling both of the task-related features (*i.e.* the synergistic and unique features) and the unimodal uncertainty noise. However, the quality and semantic richness of unimodal feature is not fully explored. In the training phase of Self-MM and MMML, we find that the prediction performance of each unimodal is quite different. Although there have been researches, such as [17; 1], that engage on estimating the reliability of unimodal prediction, it still remains to be excavated when unimodal uncertain noise is removed for better intermediate fusion. In this case, how much task-related information can unimodal features provide should be considered. In the future, we will make our effort toward this situation to overcome the issues that the quality and semantic richness of unimodal features are various.

References

- [1] Cao, B., Xia, Y., Ding, Y., Zhang, C., Hu, Q.: Predictive dynamic fusion. In: Forty-first International Conference on Machine Learning (2024), <https://openreview.net/forum?id=LYpGLrC4oq> 7
- [2] Fan, C., Zhu, K., Tao, J., Yi, G., Xue, J., Lv, Z.: Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis. IEEE Transactions on Affective Computing (2024) 4, 5
- [3] Gao, Z., Jiang, X., Xu, X., Shen, F., Li, Y., Shen, H.T.: Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26876–26885 (2024) 4, 5
- [4] Han, W., Chen, H., Poria, S.: Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 9180–9192. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). <https://doi.org/10.18653/v1/2021.emnlp-main.723>, <https://aclanthology.org/2021.emnlp-main.723/> 4
- [5] Huang, C., Chen, J., Huang, Q., Wang, S., Tu, Y., Huang, X.: Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. Information Fusion 114, 102725 (2025) 4, 5
- [6] Jaynes, E.T.: Probability theory: The logic of science. Cambridge university press (2003) 1
- [7] Kiela, D., Bhooshan, S., Firooz, H., Perez, E., Testuggine, D.: Supervised multimodal bitransformers for classifying images and text. arXiv preprint arXiv:1909.02950 (2019) 5
- [8] Li, Z., Xu, B., Zhu, C., Zhao, T.: CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. In: Carpuat, M., de Marneffe, M.C., Meza Ruiz, I.V. (eds.) Findings of the Association for Computational Linguistics: NAACL 2022. pp. 2282–2294. Association for Computational

- 189 Linguistics, Seattle, United States (Jul 2022). <https://doi.org/10.18653/v1/2022.findings-naacl.175>, <https://aclanthology.org/2022.findings-naacl.175/> 5
- 190
- 191 [9] Neuberg, L.G.: Causality: models, reasoning, and inference, by judea pearl, cambridge university press,
- 192 2000. *Econometric Theory* **19**(4), 675–685 (2003) 2
- 193 [10] Niu, T., Zhu, S., Pang, L., El Saddik, A.: Sentiment analysis on multi-view social data. In: *MultiMedia*
- 194 *Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings,*
- 195 *Part II 22*, pp. 15–27. Springer (2016) 4
- 196 [11] Tsai, Y.H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.P., Salakhutdinov, R.: Multimodal trans-
- 197 former for unaligned multimodal language sequences. In: *Proceedings of the conference. Association for*
- 198 *computational linguistics. Meeting.* vol. 2019, p. 6558. NIH Public Access (2019) 4
- 199 [12] Walters, P.: *An introduction to ergodic theory*, vol. 79. Springer Science & Business Media (2000) 3
- 200 [13] Wei, Y., Yuan, S., Yang, R., Shen, L., Li, Z., Wang, L., Chen, M.: Tackling modality heterogeneity with
- 201 multi-view calibration network for multimodal sentiment detection. In: *Proceedings of the 61st Annual*
- 202 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 5240–5252 (2023)
- 203 5
- 204 [14] Wu, Z., Gong, Z., Koo, J., Hirschberg, J.: Multimodal multi-loss fusion network for sentiment anal-
- 205 ysis. In: Duh, K., Gomez, H., Bethard, S. (eds.) *Proceedings of the 2024 Conference of the North*
- 206 *American Chapter of the Association for Computational Linguistics: Human Language Technologies*
- 207 *(Volume 1: Long Papers)*. pp. 3588–3602. Association for Computational Linguistics, Mexico City, Mex-
- 208 ico (Jun 2024). <https://doi.org/10.18653/v1/2024.naacl-long.197>, [https://aclanthology.org/2024.](https://aclanthology.org/2024.naacl-long.197/)
- 209 [naacl-long.197/](https://aclanthology.org/2024.naacl-long.197/) 5
- 210 [15] Yu, W., Xu, H., Yuan, Z., Wu, J.: Learning modality-specific representations with self-supervised multi-
- 211 task learning for multimodal sentiment analysis. In: *Proceedings of the AAAI conference on artificial*
- 212 *intelligence.* vol. 35, pp. 10790–10797 (2021) 4
- 213 [16] Zadeh, A., Chen, M., Poria, S., Cambria, E., Morency, L.P.: Tensor fusion network for multimodal
- 214 sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017) 4
- 215 [17] Zhang, Q., Wu, H., Zhang, C., Hu, Q., Fu, H., Zhou, J.T., Peng, X.: Provable dynamic fusion for low-quality
- 216 multimodal data. In: *International conference on machine learning.* pp. 41753–41769. PMLR (2023) 4, 7