

Learning in Order!

A Sequential Strategy to Learn Invariant Features for Multimodal Sentiment Analysis

Xianbing Zhao
Harbin Institute of Technology
Shezhen, China
zhaoxianbing_hitsz@163.com

Lizhen Qu
Monash University
Melbourne, Australia
lizhen.qu@monash.edu

Tao Feng
Monash University
Melbourne, Australia
taofengsd@gmail.com

Jianfei Cai
Monash University
Melbourne, Australia
asjfc@ntu.edu.sg

Buzhou Tang
Harbin Institute of Technology
Shenzhen, China
tangbuzhou@gmail.com

ABSTRACT

This work proposes a novel and simple sequential learning strategy to train models on videos and texts for multimodal sentiment analysis. To estimate sentiment polarities on unseen out-of-distribution data, we introduce a multimodal model that is trained either in a single source domain or multiple source domains using our learning strategy. This strategy starts with learning domain invariant features from text, followed by learning sparse domain-agnostic features from videos, assisted by the selected features learned in text. Our experimental results demonstrate that our model achieves significantly better performance than the state-of-the-art approaches on average in both single-source and multi-source settings. Our feature selection procedure favors the features that are independent to each other and are strongly correlated with their polarity labels. To facilitate research on this topic, the source code of this work will be publicly available upon acceptance.

KEYWORDS

MSA, OOD, Invariant Features

1 INTRODUCTION

Multimodal Sentiment Analysis (MSA) is concerned with understanding people’s attitudes or opinions based on information from more than one modalities, such as videos and texts. It finds rich applications in both industry and research communities, such as understanding spoken reviews of target products posted on YouTube and developing multimodal AI assistants for mental health support. Prior MSA approaches make an impractical assumption that training and test data comprise independent identically distributed samples [60, 66, 79, 82–84]. However, training datasets are available only for a handful of applications that satisfy that assumption. Therefore, this work aims to remove the assumption such that MSA models trained on a single domain or multiple source domains can work robustly on *unseen* out-of-distribution (OOD) data, without leveraging any target domain data.

To enable models to work robustly across domains, a key idea is to exploit domain invariant sparse representations, which serve as causes of target labels from a causal perspective [65, 70]. In contrast, spurious correlations, which do not indicate causal relations,

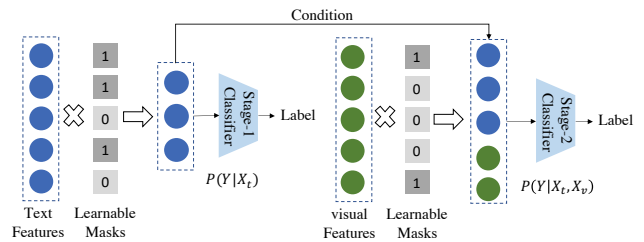


Figure 1: Classifiers employ learnable masks to identify domain-invariant text features first, conditioned on which the classifiers learn domain-invariant features from videos.

impede the generalization capability of pre-trained foundation models [6, 26]. Existing MSA models heavily rely on jointly learned multimodal features for sentiment analysis [28]. However, the spurious features of the visual modality may adversely affect the features of the text modality, leading to inaccurate prediction outcomes [24, 25, 83]. Therefore, it would be interesting to investigate i) **how to automatically identify domain invariant representations for MSA**, and ii) **what are the key characteristics of domain invariant features in a multimodal setting**.

To answer the above research questions, as illustrated in Figure 1, we propose a **Sequential Strategy to Learn Invariant Features (S²LIF)** for building a domain generalization (DG) MSA model based on videos and texts. Instead of learning domain-agnostic features simultaneously from all modalities, our technique first leverages the sparse masking technique [35] to select invariant hidden features from texts, followed by learning the invariant features from videos, conditioned on the selected textual features. *To the best of our knowledge, it is the first time to report the importance of feature learning order for domain generalization.* We conduct extensive experiments to i) demonstrate the superiority of our approach in comparison with the competitive baselines in both single source domain and multi-source domain settings, and ii) investigate key characteristics of selected features using our approach. Our key contributions are summarized as follows:

- We introduce a novel domain generalization MSA model, which explicitly learns domain-invariant features and mitigates spurious domain-specific features by adopting a sparse masking technique.
- We propose a new sequential multimodal learning strategy, which extracts the domain-invariant features from texts first, followed by employing them to identify the features relevant to labels from videos using the sparsity technique.
- We demonstrate empirically that i) our sequential multimodal learning strategy prefer selecting the domain invariant features in the visual modality, which are independent of the selected features in the text modality and strongly correlate with labels; and ii) it is important to adhere to the learning order of our approach to mitigate spurious correlations, because it is evident that the alternative learning order or learning all modalities simultaneously using the same sparsity technique leads to inferior performance.

2 RELATED WORK

2.1 Multimodal Sentiment Analysis

MSA methods can be roughly divided into two categories: 1) Multimodal Representation Learning aims to learn fine-grained multimodal representation, which provides rich decision evidence for multimodal sentiment prediction. They employ a disentangled technique to learn modality-common and modality-specific representations to mitigate the heterogeneity of multimodal representations [29, 67, 74, 77]. 2) Multimodal Fusion aims to learn cross-modal information transfer by designing complex cross-modal interactive networks. The development of multimodal fusion methods has evolved from multi-modal tensor fusion [79] to cross-modal attention [45, 46, 66, 76, 80, 82–85]. The current MSA methods only train and test on a specific domain, and do not consider the generalization ability of the model. They suffer performance degradation when tests on out-of-distribution data, so learning robust MSA models is essential.

2.2 Domain Generalization

Domain generalization aims to design a deep neural network model that learns domain-invariant features and is able to maintain stable performance in both the source domain and multiple unseen target domains. Numerous domain generalization methods have been proposed to learn domain-invariant features for single-source or multi-source domain generalization [5, 12–18, 23, 27, 31, 32, 38, 39, 41, 50, 53, 56, 61, 75]. We roughly divide current methods of domain generalization into three categories: 1) Learning invariant features aims to capture the domain-generalized features to reduce the dependence of features on specific domains and to achieve high performance on unseen domains [50]. 2) Optimize algorithm aims to learn domain-invariant features and remove domain-specific features [5, 10, 23, 39, 54, 61], such as adversarial training and meta-learning, through tailored designed network structures. 3) Data augmentation aims to generate new data to improve the generalization performance of the model, and these generated new data are out-of-distribution samples different from the source domain [69, 71, 72].

2.3 Causal Representation Learning

From the perspective of data generation, causal representation learning considers that raw data is entangled with two parts of features: correlated features with label (domain-invariant features) and spuriously correlated features with the label (domain-specific features). The goal is to disentangle domain-invariant features and domain-specific features. Domain-invariant features guarantee stable performance in different test environments [1, 51]. Based on this assumption, numerous methods attempt to learn domain invariant features [2, 3, 11, 33, 58]. Following previous work, our proposed approach aims to learn the domain-invariant features (i.e., the features correlated with the label), while removing the features domain-specific (i.e., the spuriously correlated features with the label). Concretely, we adopt sparse techniques to remove spuriously correlated features with the label [37, 44, 48, 62].

3 METHOD

Problem Statement. The goal of domain generalization for MSA is to train a deep neural network model on a single-source or multi-source domains $\mathcal{D}_S = \{\mathcal{D}_S^1, \mathcal{D}_S^2, \dots, \mathcal{D}_S^N\}$ and evaluate the model on the unseen target domains $\{\mathcal{D}_T^1, \mathcal{D}_T^2, \dots, \mathcal{D}_T^M\}$, where \mathcal{D} denotes a dataset in a domain, M and N denote the number of source domains and target domains, respectively. We consider each MSA task as a k -ways classification task. The dataset in a source domain is denoted by $\mathcal{D}_S = \{\mathcal{X}_{\{t,v\}}^i, y_i\}_{i=1}^n$, where $\mathcal{X}_{\{t,v\}}^i \in \mathbb{R}^d$, $y \in \mathbb{R}^K$, while \mathcal{D}_T is a dataset in a target domain. The goal is to learn multimodal domain-invariant features for sentiment polarity prediction in unseen domains without using target domain data for training.

Model Overview. Our work is motivated by the functional lottery ticket hypothesis [42] suggesting that there is a subnetwork that can achieve better out-of-distribution performance than the original network. Hence, We employ the sparse masking techniques to identify a subset of hidden features in the multimodal setting. The findings of our empirical studies indicate the importance of the learning order between modalities for domain generalization performance.

The architecture of our model is illustrated in Figure 2. Given a text and a sequence of video frames $\mathcal{X}_{\{t,v\}}$, we employ a pre-trained encoders ELECTRA [20] and VGGFace2 with a 1-layer Transformer encoder [22] to map them to respective hidden representations x_t and x_v . To achieve sparsity in hidden representations, our model generates a mask vector $m_{\{t,v\}}$ with the mask function f_{mask} to select domain-invariant features $x_{\{t,v\}}^c$ from $\mathcal{X}_{\{t,v\}}$. The mask function is characterized by the learnable parameter $r_{\{t,v\}}$ and threshold $s_{\{t,v\}}$. The feature selection in a modality is achieved by computing the dot-product between the mask vectors and the corresponding hidden representations. We empirically find that text is the superior modality in comparison with videos based on their performance in each modality. Our further studies show that conditioning on the strong text features reduces the selection of visual features that correlate with those text features. On the one hand, reduction of statistical dependencies between features leads to improvement of generalization performance. On the other hand, selection of features adhere to the functional lottery ticket hypothesis. Therefore, our text classifier g_t first selects the key features using the masking

technique, followed by learning sparse video representations for the visual classifier g_v to predict sentiment polarity conditioned on the selected text features. In addition, our model leverages prior information from video frames to eliminate redundant frames.

Keyframe-aware Masking. Given that there is a large amount of frames in a video clip, which contains redundant information [48, 62]. The frame sequence \bar{x}_v of a video clip contains rich priors, which explicitly correspond to neighboring frames. We can easily obtain the motion of the video frame sequence to guide the masking of redundant frames according to the temporal difference. We employ global and local neighbor frames to select informative frames x_v and constrain the semantic invariance of video frames by reconstructing losses \mathcal{L}_{recon} . Note that this part is not our main contribution, and it is an extension of previous work [48]. See supplementary materials for details.

Sequential Multimodal Learning. The selection of domain invariant features is also motivated from a causal perspective. The logit of the classifier is computed as the product between the features x and the weights W of the label k from the classification layer g .

$$O_k = W_{\{k\}}^T \cdot x = \sum_{* \in \{t, v\}} \sum_{j=1}^{d_t} W_{*(j,k)} \cdot x_{*j}, * \in \{t, v\}, \quad (1)$$

where the subscripts j and k denote j -th feature and k -th class respectively. For all polarity labels with both modalities, we obtain a matrix R as follows, where each element $w_{*(j,k)} \cdot x_{*j}, * \in \{t, v\}$ represents the evidence of the classifier.

$$R = \begin{bmatrix} w_{t(1,1)} x_{t1} & \dots & w_{t(1,K)} x_{t1} \\ w_{t(2,1)} x_{t2} & \dots & w_{t(2,K)} x_{t2} \\ \vdots & \ddots & \vdots \\ w_{t(d_t,1)} x_{td_t} & \dots & w_{t(d_t,K)} x_{td_t} \\ w_{v(1,1)} x_{v1} & \dots & w_{v(1,K)} x_{v1} \\ w_{v(2,1)} x_{v2} & \dots & w_{v(2,K)} x_{v2} \\ \vdots & \ddots & \vdots \\ w_{v(d_v,1)} x_{vd_v} & \dots & w_{v(d_v,K)} x_{vd_v} \end{bmatrix} \quad (2)$$

By analyzing the matrix R , we conclude that **1) Y is the result of feature x estimated via a classifier.** From the causal perspective, the selected features can be seen as the causes of Y subjecting to independent noise [30, 52, 64]:

$$Y = g(Pa(Y)) + \epsilon \quad (3)$$

where the notation $Pa(Y)$ denotes the features of direct causal effects with Y , where $Pa(Y)$ is a subset of x . The function g represents the classifier. The multimodal features x are divided into two subsets, domain-specific features x^s (spurious correlated features with the label across domain) and domain-invariant features x^c (correlated features with the label across domain) [57]. We use three features $\{x_1, x_2, x_3\}$ to explain the causal relationship between x and Y . As shown in Figure 3 (a), the outcome Y is specified as $Y = g(x_1, x_3) + \epsilon, \{x_1, x_3\} \subseteq x^c$. The feature x_3 is the subset of x^s . There exist two distinct relationships between the feature sets x^s and x^c : a) there is no direct causal relationship between x_3 and

x_1 . b) there is a direct causal relationship between x_3 and x_2 . We remove the edge between x_2 and x_3 to eliminate the impact of x_3 on x_2 . Therefore, our goal is to identify the features x^c and remove the features x^s . Formally, we expect

$$\mathbb{P}(Y|do(x_i^c, x_k^s)) \neq \mathbb{P}(Y|do(x_j^c, x_k^s)) \quad (4)$$

where the features $\{x_i^c, x_j^c\} \subseteq x^c$ are selected mutual independent domain-invariant features [59]. We design learnable masks m and learnable threshold s in Section 3 to set the values of domain-specific features in x_k^s to 0. Removing the features $x_k^s \subseteq x^s$ eliminates its direct causal effects on (x_i^c, x_j^c) and the outcome Y . **2) simultaneously optimizing such entangled features $x = \{x^c, x^s\}$ for both text and visual modalities (i.e., imbalanced multimodal features) poses a special challenge for the classifier** [24, 25].

Our sequential learning strategy is also motivated by curriculum learning [8, 47, 87] that we learn the features first, which perform well on the target tasks, followed by more challenging ones.

By analyzing the causal relationship and removing spurious correlation features using multimodal learnable masks, we can obtain a new evidence matrix R^M . The form of the new evidence matrix R^M for the classifier is as follows:

$$R^M = \begin{bmatrix} w_{t(1,1)} x_{t1} m_{t1} & \dots & w_{t(1,K)} x_{t1} m_{t1} \\ w_{t(2,1)} x_{t2} m_{t2} & \dots & w_{t(2,K)} x_{t2} m_{t2} \\ \cancel{w_{t(3,1)} x_{t3} m_{t3}} & \dots & \cancel{w_{t(3,K)} x_{t3} m_{t3}} \\ \vdots & \ddots & \vdots \\ w_{t(d_t,1)} x_{td_t} m_{td_t} & \dots & w_{t(d_t,K)} x_{td_t} m_{td_t} \\ w_{v(1,1)} x_{v1} m_{v1} & \dots & w_{v(1,K)} x_{v1} m_{v1} \\ w_{v(2,1)} x_{v2} m_{v2} & \dots & w_{v(2,K)} x_{v2} m_{v2} \\ \cancel{w_{v(3,1)} x_{v3} m_{v3}} & \dots & \cancel{w_{v(3,K)} x_{v3} m_{v3}} \\ \vdots & \ddots & \vdots \\ w_{v(d_v,1)} x_{vd_v} m_{vd_v} & \dots & w_{v(d_v,K)} x_{vd_v} m_{vd_v} \end{bmatrix} \quad (5)$$

where $\{m_t, m_v\}$ denotes mask vector in Section 3. The red notation m_{t_i} and m_{v_j} represent learnable mask to select domain-invariant feature with two stages in the above equations. By analyzing the evidence matrix R^M of the classifier g and the direct causal effect with outcome Y , we can utilize the learnable mask and threshold to sequential select domain-invariant features x^c and remove domain-specific features x^s .

Multimodal Learnable Masks. Regarding how to automatically identify domain invariant representations for MSA, we design multimodal learnable masks to select features. Specifically, to remove domain-specific features, we tailor a function, denoted as f_{mask} . The inputs of f_{mask} consists of the features x from a modality, a learnable parameter r , and a dynamic threshold s . The output is domain-invariant features x^c .

$$x^c = f_{mask}(x, r, s) \quad (6)$$

where we apply the mask vector $m \in \mathbb{R}^d$ (consisting of zero and non-zero value) on the feature $x \in \mathbb{R}^d$. The mask vector m is obtained by utilizing a trainable pruning threshold $s \in \mathbb{R}^d$ and a learnable parameter $r \in \mathbb{R}^d$. Given a set of features x , our method can dynamically select features using mask vector m . We utilize the unit step function $\mathcal{F}(\cdot)$ to produce mask vector, which takes

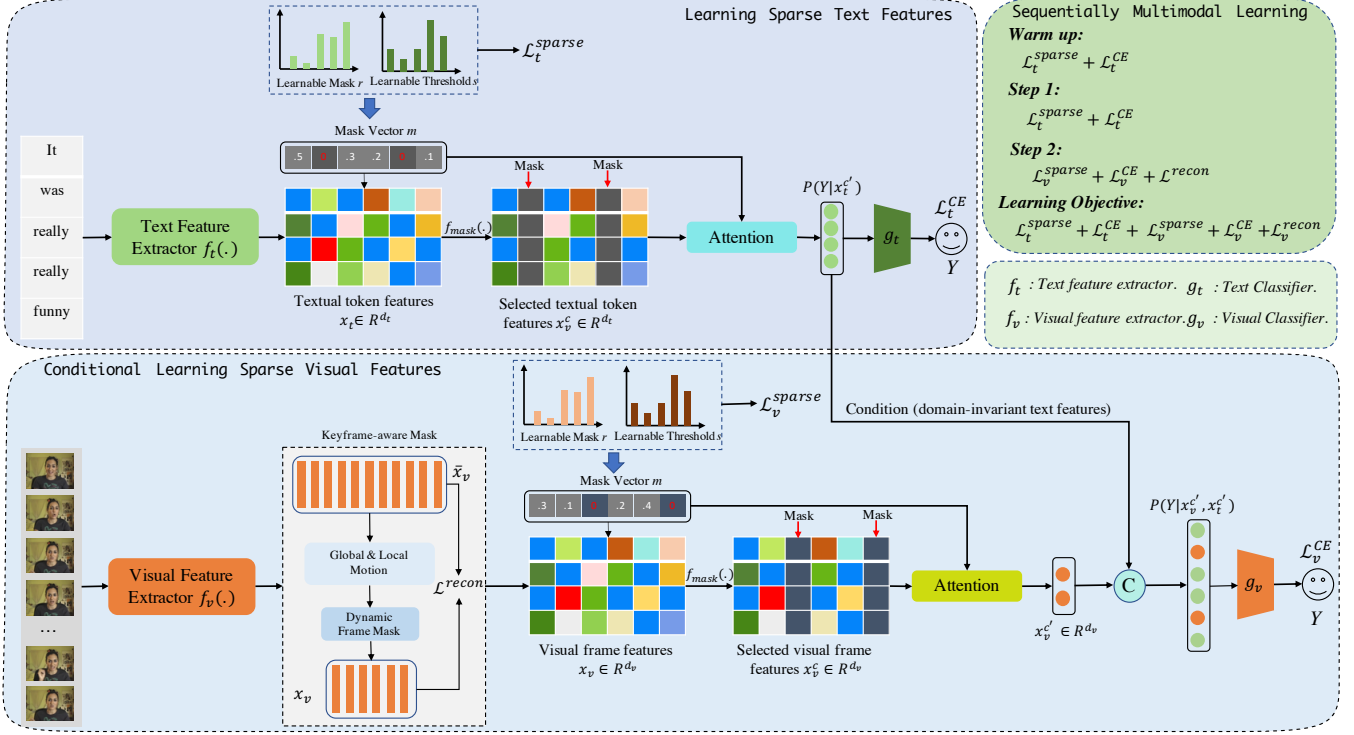


Figure 2: An overview of our proposed framework.

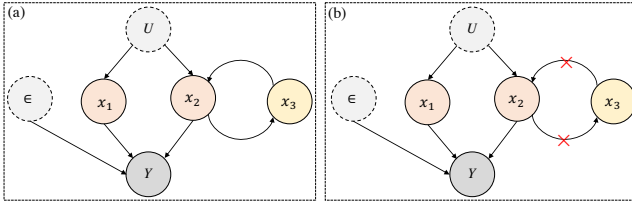


Figure 3: (a): The causal structure of the data generation process involves direct causal effects from x_1 and x_2 to Y . There exists a causal relationship between x_2 and x_3 . ϵ represents independent noise. The latent variable U serves as a confounder for x_1 and x_3 . (b) Severing the edge between x_2 and x_3 and eliminating the causal relationship.

the learnable parameters r and thresholds s as input and output the binary masks p . Formally,

$$\mathcal{F}(t) = \begin{cases} 0 & \text{if } t < 0 \\ 1 & \text{if } t \geq 0 \end{cases}, \quad (7)$$

where the binary mask p and mask vector m are obtained by:

$$p = \mathcal{F}(|r| - s), \quad (8)$$

$$m = r \odot p \quad (9)$$

$$x^c = x \odot m, \quad (10)$$

where x^c represents domain-invariant features, which remove spuriously correlated features and retain the correlated features with

the label y in training stage. It was unable to complete end-to-end training during model training. The reason is that the binary mask produced by our unit step function is non-differentiable. To overcome this issue, previous works [34, 63, 86] based on straight-through estimator (STE) [7] to estimate derivatives and design binarization function that can be back-propagation. [73] give more approximate estimates than STE to handle non-differentiable scenarios. Using this derivative estimate to approximate the unit step function allows the model to train end-to-end.

$$\frac{d}{dt} \mathcal{F}(t) = \begin{cases} 2 - 4|t|, & -0.4 \leq t \leq 0.4 \\ 0.4, & 0.4 \leq |t| \leq 1 \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

To encourage the model to learn sparse features, we add a sparse regularization term [35] to the threshold as one of the training objectives. Formally,

$$\mathcal{L}_*^{sparse} = \sum_{i=1}^N \exp(-s_i^*), \quad * \in \{t, v\}, \quad (12)$$

where the regular term $\exp(-s_i^*)$ raises the value of the dynamic threshold s , so that a few feature values can exceed the threshold to learn more sparse features. We utilize the function f_{mask} to obtain domain-invariant features of textual and visual tokens. Formally,

$$x_*^c = f_{mask}(x_*, r_*, s_*), \quad * \in \{t, v\} \quad (13)$$

where the definitions of f_{mask} , r_* , and s_* are specified in Equation 6 ~ 10.

Apart from learning the domain-invariant features of each token, we also calculate the similarity between each token and the learnable mask to learn domain-invariant tokens. Formally,

$$a^{*c_j} = \text{sim}(m_*, x_*^{c_j}), * \in \{t, v\}, \quad (14)$$

$$x_*^{c'} = \sum_{j=1}^{\tau_*} x_*^{c_j T} \cdot a^{*c_j}, * \in \{t, v\}, \quad (15)$$

where $x_*^{c_j}$ denotes j -th token of text and visual modalities. The $x_t^{c'}$ and $x_v^{c'}$ represent the features of fused domain-invariant tokens. The symbol sim denotes similarity. The symbol a denotes attention weight. The classifier g_t and g_v takes inputs $x_t^{c'}$ and $x_v^{c'}$, and outputs logits O_t and logits O_v . Formally,

$$O_t = g_t(x_t^{c'}); O_v = g_v([x_t^{c'}; x_v^{c'}]) \quad (16)$$

where ';' denotes concatenation along the feature dimension.

Learning Objective. Initially, we employ the classifier to learn domain-invariant features from the text modality (i.e. text modality). Formally,

$$\mathcal{L}_t = \mathcal{L}_t^{CE} + \alpha \cdot \mathcal{L}_t^{sparse} \quad (17)$$

Subsequently, we utilize the domain-invariant features from the text modality to assist in selecting domain-invariant features from the visual modality. Formally,

$$\mathcal{L}_v = \mathcal{L}_v^{CE} + \alpha \cdot \mathcal{L}_v^{sparse} + \mathcal{L}_{recon} \quad (18)$$

where the symbol \mathcal{L}_*^{CE} denotes *Cross-Entropy* loss and α is hyper-parameter. Accordingly, the overall learning objective is:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_v \quad (19)$$

4 EXPERIMENTS

4.1 Datasets

We select three typical MSA benchmark datasets: CMU-MOSI [81], CMU-MOSEI [82] and MELD [55]. The detailed partition of the dataset is included in the supplementary materials.

4.2 Implementation Detail

We employ text pre-trained language model Electra [20] and visual pre-trained model VGG Face2 [9], extracting features from both textual content and video frames. We use a multilayer perceptron to unify the multimodal feature dimensions and a 1-layer Transformer encoder [22] to model the multimodal data of the sequence. The batch size and epoch are set to 16 and 200, and the learning rate is configured to $7e-5$. Warm up epoch is 3. Our implementation is executed using the PyTorch framework with Adam optimizer [36] on the V100 GPU.

4.3 Baselines

We select the state-of-the-art model in the field of MSA, MLLM and DG (OOD) as the Baseline.

MULT [67] designs a Multimodal Transformer to align multi-modal sequential data and capture cross-modal information interaction.

ALMT [83] employs non-verbal modalities to reinforce the features of the text modality several times and dismisses the non-verbal information after completing the reinforcement process.

MAD [56] designs a two-stage learning strategy, learning domain-specific and domain-invariant features respectively to constrain the two features by regular terms.

RIDG [17] aligns the labels of each class with the classified evidence to ensure the domain generalization of the model.

MLLM. We selected five multimodal large language models, including Blip-2 [40], InstructBlip [21], [43] and Qwen-VL [4] with excellent performance from the benchmark [78] of the multimodal large model as the baseline.

4.4 Evaluation Criteria

The distribution of the dataset is approximately balanced. We evaluate the model performance using a 3-class accuracy metric, specifically [Positive, Neutral, Negative].

4.5 Results and Discussions

Overall Comparisons. To justify the effectiveness of our proposed S^2LIF model, we compared the model with the following state-of-the-art baseline in the field of MSA and DG. Models that focus on capturing cross-modal dependencies, called MULT and ALMT. Models that aims to learn domain-invariant features, namely, MAD and RIDG. Tables 1 and 2 show the results of the comparison. By analyzing these two tables, we draw the following conclusions: **i)** The MSA method shows visual performance in the unseen domain. With the addition of our multimodal learnable masks, the traditional models also gain the ability of DG. The fact demonstrates the effectiveness of sparse mask in DG. **ii)** Our model significantly outperforms the multimodal large model in 4 of the 6 Settings. We speculate that there is contamination from emotional datasets during the training phase of the multimodal large language model. In the two settings with better performance, the logits of InstructBlip for correctly predicted samples exceed 0.93, significantly higher than the logits generated by other multimodal large language models, which are around 0.65. **iii)** The model performance in sequential multimodal learning is better than that in non-sequential multimodal learning when we distinguish text and visual modalities. This demonstrates the effectiveness of the sequential multimodal learning strategy.

Existence of Domain-invariant Features. An essential assumption in our study is the presence of domain-invariant features in cross-domain multimodal data. To gain insight into this assumption, we visualized the selected and removed features for each domain using a heatmap. We marked positions with '1' where the features are consistently selected across domains. From Figure 4, we could conclude that there is a presence of domain-invariant features across multiple domains, and our proposed model can automatically select the domain-invariant features. Moreover, we visualized the proportion of features retained during the training phase. Figure 5 illustrates the proportion of features retained for both the text and visual modalities during the training phase.

Cross-modal Feature Correlation Analysis. Apart from the superior performance, the key advantage of our proposed model compared to other models is that its sequential multimodal learning. It can conditionally assist visual modalities in selecting domain-invariant features based on the domain-invariant features learned

Table 1: The performance (accuracy of 3-classification) of single-source domain generalization. The symbols V, T , and M denote using visual, textual, and multimodal features, respectively. The symbols $T \rightarrow V$ and $V \rightarrow T$ indicate the multimodal learning order. $T&V$ denotes simultaneous learning. 'Frozen' and 'Fine tuning' represents freezing and fine-tuning the parameter of pre-trained language model. We train the model on the source domain and infer on both the source and target domains.

| Category | Method | Single-source Setting A | | | Single-source Setting B | | | Single-source Setting C | | |
|---------------------------------|--|-------------------------|---------------|--------------|-------------------------|---------------|--------------|-------------------------|---------------|--------------|
| | | Source Domain | Target Domain | | Source Domain | Target Domain | | Source Domain | Target Domain | |
| | | MOSEI | MOSI | MELD | MOSI | MELD | MOSEI | MELD | MOSI | MOSEI |
| MSA | MuLT (M-Frozen) (ACL2019) | 0.644 | 0.693 | 0.516 | 0.609 | 0.344 | 0.452 | 0.663 | 0.258 | 0.435 |
| | MuLT (M-Fine tuning) (ACL2019) | 0.691 | 0.740 | 0.526 | 0.736 | 0.400 | 0.506 | 0.687 | 0.453 | 0.493 |
| | ALMT (M-Frozen) (EMNLP2023) | 0.611 | 0.688 | 0.468 | 0.548 | 0.373 | 0.465 | 0.686 | 0.306 | 0.457 |
| | ALMT (M-Fine tuning)(EMNLP2023) | 0.676 | 0.675 | 0.540 | 0.760 | 0.522 | 0.513 | 0.679 | 0.478 | 0.440 |
| MSA+Mask | MuLT + Mask (M-Frozen) (ACL2019) | 0.649 | 0.710 | 0.545 | 0.625 | 0.435 | 0.487 | 0.667 | 0.325 | 0.456 |
| | MuLT + Mask (M-Fine tuning) (ACL2019) | 0.693 | 0.759 | 0.554 | 0.766 | 0.500 | 0.553 | 0.706 | 0.519 | 0.510 |
| | ALMT + Mask (M-Frozen) (EMNLP2023) | 0.642 | 0.693 | 0.509 | 0.574 | 0.472 | 0.473 | 0.697 | 0.376 | 0.460 |
| | ALMT + Mask (M-Fine tuning) (EMNLP2023) | 0.687 | 0.749 | 0.562 | 0.777 | 0.541 | 0.586 | 0.700 | 0.553 | 0.497 |
| OOD | MAD (V) (CVPR2023) | 0.450 | 0.279 | 0.291 | 0.390 | 0.214 | 0.312 | 0.468 | 0.218 | 0.326 |
| | MAD (T-Frozen) (CVPR2023) | 0.413 | 0.172 | 0.349 | 0.344 | 0.205 | 0.334 | 0.482 | 0.154 | 0.346 |
| | MAD (T-Fine tuning)(CVPR2023) | 0.464 | 0.154 | 0.331 | 0.491 | 0.204 | 0.296 | 0.660 | 0.157 | 0.311 |
| | MAD (M-Frozen) (CVPR2023) | 0.448 | 0.304 | 0.423 | 0.374 | 0.303 | 0.346 | 0.495 | 0.243 | 0.368 |
| | MAD (M-finetune) (CVPR2023) | 0.670 | 0.744 | 0.527 | 0.746 | 0.505 | 0.563 | 0.683 | 0.419 | 0.484 |
| | RIDG (V) (ICCV2023) | 0.317 | 0.313 | 0.419 | 0.437 | 0.221 | 0.335 | 0.471 | 0.262 | 0.358 |
| | RIDG (T-Frozen) (ICCV2023) | 0.555 | 0.384 | 0.477 | 0.481 | 0.256 | 0.351 | 0.491 | 0.311 | 0.367 |
| | RIDG (T-Fine tuning) (ICCV2023) | 0.665 | 0.728 | 0.489 | 0.644 | 0.527 | 0.487 | 0.657 | 0.495 | 0.417 |
| | RIDG (M-Frozen) (ICCV2023) | 0.572 | 0.467 | 0.520 | 0.505 | 0.318 | 0.392 | 0.657 | 0.319 | 0.425 |
| RIDG (M-Fine tuning) (ICCV2023) | 0.657 | 0.736 | 0.523 | 0.695 | 0.540 | 0.583 | 0.680 | 0.513 | 0.501 | |
| MLLM | Blip-2 (ICML2023) | 0.397 | 0.290 | 0.448 | 0.290 | 0.448 | 0.397 | 0.448 | 0.290 | 0.397 |
| | InstructBlip (NeurIPS2024) | 0.540 | 0.739 | 0.492 | 0.739 | 0.492 | 0.540 | 0.492 | 0.739 | 0.540 |
| | LLava-1.5-7B (NeurIPS2023) | 0.510 | 0.351 | 0.527 | 0.351 | 0.527 | 0.510 | 0.527 | 0.351 | 0.510 |
| | LLava-1.5-13B (NeurIPS2023) | 0.453 | 0.192 | 0.496 | 0.192 | 0.496 | 0.453 | 0.496 | 0.192 | 0.453 |
| | Qwen-VL | 0.455 | 0.250 | 0.536 | 0.250 | 0.536 | 0.455 | 0.536 | 0.250 | 0.455 |
| Ours | $S^2LIF V \rightarrow T$ (M-Frozen) | 0.653 | 0.718 | 0.513 | 0.631 | 0.421 | 0.492 | 0.645 | 0.383 | 0.445 |
| | $S^2LIF T&V$ (M-Frozen) | 0.651 | 0.720 | 0.535 | 0.629 | 0.450 | 0.504 | 0.658 | 0.380 | 0.464 |
| | $S^2LIF T \rightarrow V$ (M-Frozen) | 0.660 | 0.745 | 0.543 | 0.638 | 0.465 | 0.517 | 0.643 | 0.408 | 0.488 |
| | $S^2LIF V \rightarrow T$ (M-Fine tuning) | 0.688 | 0.759 | 0.539 | 0.775 | 0.498 | 0.606 | 0.683 | 0.529 | 0.491 |
| | $S^2LIF T&V$ (M-Fine tuning) | 0.689 | 0.755 | 0.540 | 0.742 | 0.524 | 0.584 | 0.677 | 0.481 | 0.504 |
| | $S^2LIF T \rightarrow V$ (M-Fine tuning) | 0.701 | 0.774 | 0.572 | 0.762 | 0.556 | 0.613 | 0.692 | 0.580 | 0.519 |

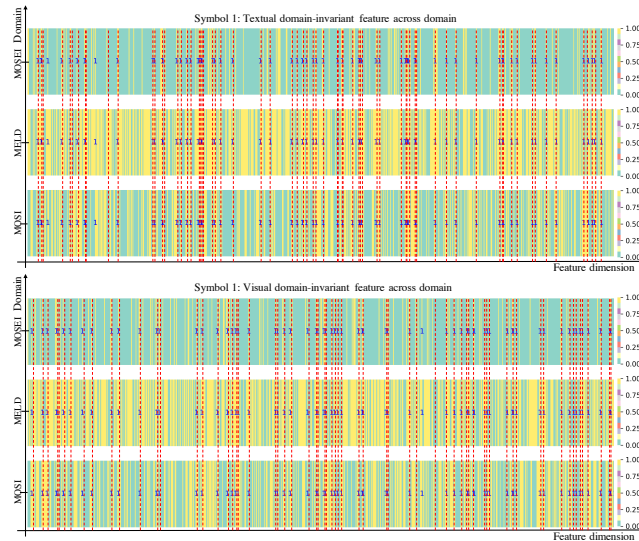


Figure 4: Visualization of domain-invariant features across domain.

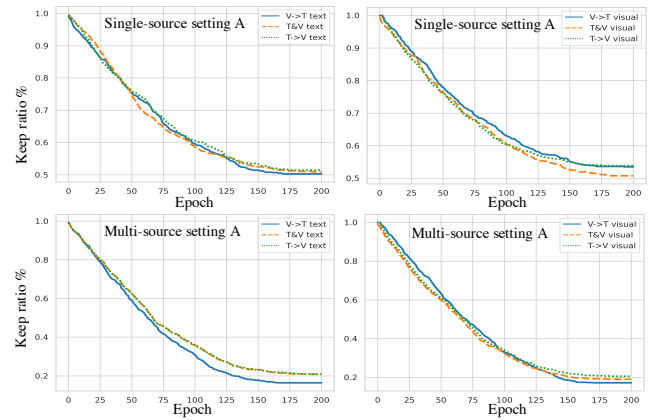


Figure 5: The proportion of domain-invariant features.

from the text modality. The features of these visual modalities prefer mutually independent from the features of the text modality, allowing the information learned from the visual modalities to complement that of the text modality. For each domain-invariant feature $x_{t_i}^{C'} \subset x_t^{C'}$ from the text modality, we employed Fisher's

Table 2: The performance (accuracy of 3-classification) of multi-source domain generalization.

| Category | Method | Multi-source Setting A | | Multi-source Setting B | | Multi-source Setting C | |
|---------------------------------|---------------------------------------|------------------------|---------------|------------------------|---------------|------------------------|---------------|
| | | Source Domain | Target Domain | Source Domain | Target Domain | Source Domain | Target Domain |
| | | MOSEI/MELD | MOSI | MOSI/MELD | MOSEI | MOSI/MOSEI | MELD |
| MSA | MuLT (M-Frozen) (ACL2019) | 0.619/0.625 | 0.621 | 0.666/0.646 | 0.470 | 0.676/0.636 | 0.464 |
| | MuLT (M-Finetune) (ACL2019) | 0.674/0.710 | 0.660 | 0.742/0.673 | 0.549 | 0.797 /0.661 | 0.494 |
| | ALMT (M-Frozen) (EMNLP2023) | 0.630/0.669 | 0.597 | 0.660/0.685 | 0.454 | 0.664/0.591 | 0.471 |
| | ALMT (M-Finetune) (EMNLP2023) | 0.683/0.697 | 0.654 | 0.753/ 0.770 | 0.528 | 0.746/0.680 | 0.516 |
| MSA+Mask | MuLT (M-Frozen) (ACL2020) | 0.647/0.659 | 0.645 | 0.644/0.688 | 0.505 | 0.730/0.654 | 0.531 |
| | MuLT (M-Finetune) (ACL2020) | 0.682/0.708 | 0.683 | 0.769/0.721 | 0.576 | 0.765/0.682 | 0.561 |
| | ALMT (M-Frozen) (EMNLP2023) | 0.640/0.657 | 0.640 | 0.645/0.682 | 0.512 | 0.673/0.631 | 0.527 |
| | ALMT (M-Finetune) (EMNLP2023) | 0.684/ 0.711 | 0.681 | 0.771 /0.721 | 0.570 | 0.787/0.688 | 0.566 |
| OOD | MAD (V) (CVPR2023) | 0.437/0.468 | 0.306 | 0.365/0.411 | 0.306 | 0.355/0.436 | 0.356 |
| | MAD (T-frozen) (CVPR2023) | 0.404/0.481 | 0.306 | 0.393/0.444 | 0.319 | 0.349/0.365 | 0.388 |
| | MAD (T-finetune) (CVPR2023) | 0.444/0.691 | 0.274 | 0.432/0.653 | 0.275 | 0.438/0.481 | 0.364 |
| | MAD (M-Finetune) (CVPR2023) | 0.485/0.672 | 0.297 | 0.484/0.676 | 0.305 | 0.445/0.511 | 0.383 |
| | MAD (M-Frozen) (CVPR2023) | 0.431/0.480 | 0.316 | 0.370/0.445 | 0.349 | 0.371/0.349 | 0.428 |
| | RIDG (V) (ICCV2023) | 0.410/0.332 | 0.355 | 0.339/0.199 | 0.367 | 0.154/0.411 | 0.381 |
| | RIDG (T-frozen) (ICCV2023) | 0.548/0.623 | 0.422 | 0.561/0.643 | 0.440 | 0.571/0.552 | 0.407 |
| | RIDG (T-finetune) (ICCV2023) | 0.646/0.656 | 0.635 | 0.737/0.666 | 0.556 | 0.752/0.663 | 0.499 |
| | RIDG (M-Frozen) (ICCV2023) | 0.550/0.630 | 0.486 | 0.605/0.654 | 0.465 | 0.603/0.594 | 0.445 |
| RIDG (M-Fine tuning) (ICCV2023) | 0.659/0.678 | 0.645 | 0.747/0.674 | 0.555 | 0.766/0.672 | 0.527 | |
| MLLM | Blip-2 (ICML2023) | 0.397/0.448 | 0.290 | 0.290/0.448 | 0.397 | 0.290/0.397 | 0.448 |
| | InstructBlip (NeurIPS2024) | 0.540/0.492 | 0.739 | 0.739/0.492 | 0.540 | 0.739/0.540 | 0.492 |
| | LLava-1.5-7B (NeurIPS2023) | 0.510/0.527 | 0.351 | 0.351/0.527 | 0.510 | 0.351/0.510 | 0.527 |
| | LLava-1.5-13B (NeurIPS2023) | 0.453/0.496 | 0.192 | 0.192/0.496 | 0.453 | 0.192/0.453 | 0.496 |
| | Qwen-VL | 0.455/0.536 | 0.250 | 0.536/0.455 | 0.250 | 0.250/0.455 | 0.536 |
| Ours | $S^2LIF V \rightarrow T$ (M-Frozen) | 0.660/0.696 | 0.658 | 0.626/0.678 | 0.484 | 0.740/0.655 | 0.493 |
| | $S^2LIF V \& T$ (M-Frozen) | 0.657/0.700 | 0.659 | 0.653/0.676 | 0.487 | 0.702/0.649 | 0.505 |
| | $S^2LIF T \rightarrow V$ (M-Frozen) | 0.650/0.699 | 0.674 | 0.625/0.679 | 0.529 | 0.737/0.638 | 0.532 |
| | $S^2LIF V \rightarrow T$ (M-Finetune) | 0.679/0.712 | 0.677 | 0.758/0.707 | 0.566 | 0.778/ 0.691 | 0.557 |
| | $S^2LIF V \& T$ (M-Finetune) | 0.686/0.706 | 0.686 | 0.762/0.704 | 0.539 | 0.768/0.671 | 0.546 |
| | $S^2LIF T \rightarrow V$ (M-Finetune) | 0.687 /0.710 | 0.687 | 0.759/0.723 | 0.581 | 0.791/0.687 | 0.578 |

z-test to calculate the ratio of features in the domain-invariant feature set $x_j^{c'}$ of the visual modality that are independent and dependent of that specific feature $x_{t_i}^{c'}$. From Figure 6, we could see that, conditioning on the text modality, the model exhibits a higher proportion of independence among domain-invariant features across modalities. These results demonstrate that our proposed sequential multimodal learning strategy in Equation (4) is capable of learning more effective, sparse, and independent cross-modal features.

Intra-modal Feature Correlation Analysis. To valid the independence among the learned domain-invariant features, we conducted Fisher’s z-test on the features in the Multi-source setting A with intra-modality. Specially, we selected $x_j^{c'}$ from the domain-invariant feature set $x^{c'}$ and computed the ratio of features in the set that are independent and dependent of $x_j^{c'}$. From Figure 7, we could be observed that, for the learned domain-invariant feature set, the proportion of features that are independent of any other feature in the set is significantly higher than the proportion of features that are dependent. This observation substantiates our assumption in Equation (4) that the combination of the learnable mask and classifier can effectively learn sparse and independent features.

Correlation Analysis Between Features and Label. To demonstrate the effectiveness of sequential multimodal learning, we also employed Fisher’s z-test to analyze the correlation between the learned domain-invariant features and labels. From Figure 8, we could observe that sequential multimodal learning is capable of capturing

Table 3: Ablation study on Multi-source Setting A.

| Method | Multi-source Setting A | |
|-------------------------------------|------------------------|---------------|
| | Source Domain | Target Domain |
| | MOSEI/MELD | MOSI |
| Add Noise | 0.367/0.199 | 0.339 |
| Using DS | 0.372/0.202 | 0.341 |
| w/o key-frame mask | 0.642/0.695 | 663 |
| $S^2LIF T \rightarrow V$ (M-Frozen) | 0.650/0.699 | 0.674 |

more features that are dependent with labels. The removed features exhibit independent with the labels. This experimental result validates the efficacy of sparse masks for feature selection and the effectiveness of sequential multimodal learning (as described in Equation (3) and (4)).

Ablation Studies. To gain the insights into our sequential multimodal learning strategy, We compare our model with the following variants: 1) Reordering sequence learning, including $T \rightarrow V$, $V \rightarrow T$, $T \& V$, where they respectively denote sequential multimodal learning with textual modality as the condition, with visual modality as the condition, and simultaneous learning of textual and visual modality. 2) **Add Noise**, introducing noise by replacing domain-invariant features with random noise as evidence for the

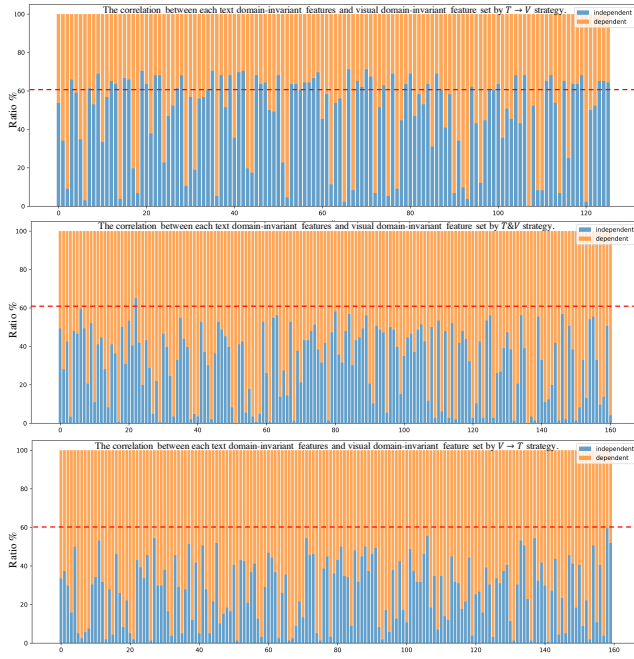


Figure 6: X-axis: Single textual domain-invariant feature. Y-axis: The independent and dependent ratio of the visual domain-invariant feature set to the each textual domain-invariant feature.

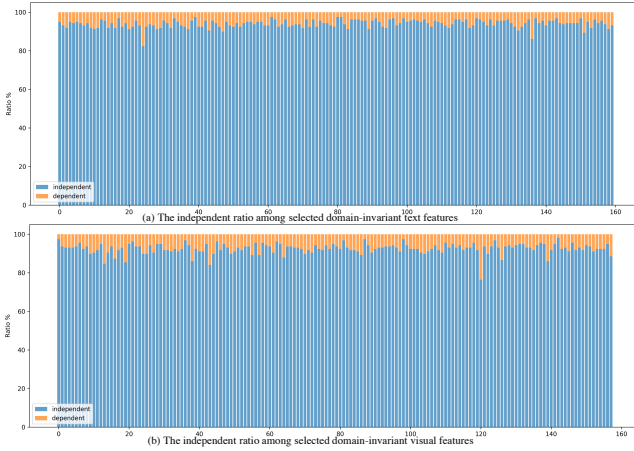


Figure 7: X-axis: Single domain-invariant feature of intra-modality. Y-axis: The independent and dependent ratio between single domain-invariant feature and the other domain-invariant features of intra-modality.

classifier. 3) **Using DS**, utilizing domain-specific features as evidence for the classifier. 4) **w/o key-frame mask**, eliminating the key-frame masking module.

From Table 1, 2 and 3, we could see that leveraging the text modality as a condition yields higher performance. Table 2 reveals that replacing the learned domain-invariant features with noise results

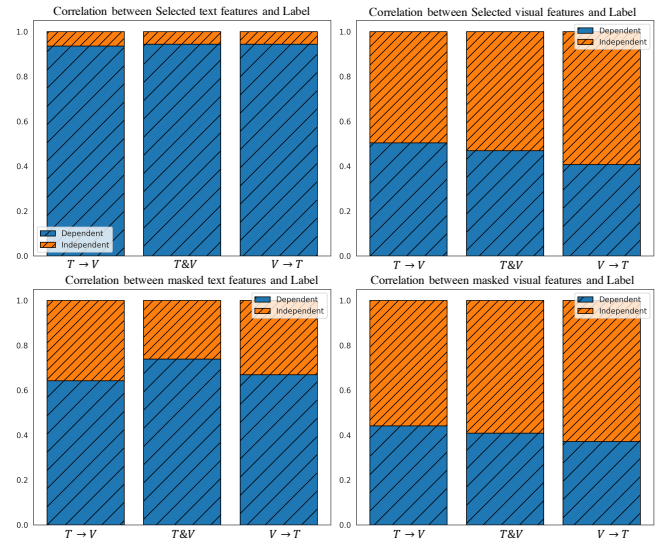


Figure 8: The correlated proportion of domain-invariant and domain-specific features with label.


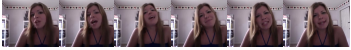

| Modal Content | Ground Truth | Prediction |
|--|--------------|------------|
|  | Negative | Negative ✓ |
|  | Neutral | Neutral ✓ |
|  | Positive | Positive ✓ |

Figure 9: The predictions on the testset of Multi-source Setting A.

in a modest performance decline, with domain-specific features outperforming random noise to a slight extent. These observations reflect the following key insights: 1) The effectiveness of sequential multimodal learning. 2) The capability of our model to efficiently learn domain-invariant features. 3) Our model effectively eliminates domain-specific features that do not contribute significantly to classification.

4.6 Case Study.

To qualitatively validate the effectiveness of our proposed model, we showcase the predictive outcomes of our model on several samples, encompassing positive, negative, and neutral sentiments. As shown in Figure 9, our model demonstrates accurate recognition of all three sentiment polarity. This indicates the robustness of our model on unseen domains.

5 CONCLUSION

In this paper, we design a sequential multimodal learning strategy to learn cross-domain invariant features for MSA. Specifically, we

first employ learnable masks and classifiers to learn the invariant features from texts, and then select the invariant features of videos, conditioned on the selected text features. The experiment demonstrates the efficacy of our model in both single-domain and multi-source domain settings. Based on extensive experiments, we conclude that i) the learning order between modalities is important for domain generalization performance, and ii) our learning strategy prefers the selection of features that are statistically independent to each other, in particular between modalities.

In the future, we will consider including more modalities, such as audio modality, to analyze the correlation between cross-modal invariant features in cross-domain scenarios.

A METHODOLOGY

A.1 Keyframe-aware Masking

Given that there is a large amount of frames in a video clip, which contains redundant information. The frame sequence \bar{x}_v of a video clip contains rich priors, which explicitly correspond to neighboring frames. We can easily obtain the motion of the video frame sequence to guide the masking of redundant frames according to the temporal difference. Temporal neighbor frames in a video clip can be divided into global neighbor frames and local neighbor frames. The local and global difference information are defined as:

$$M_i^{local} = \frac{1}{2k} \left(\sum_{j=i-k}^i \bar{x}_{v_j} + \sum_{j=i+1}^{i+k} \bar{x}_{v_j} \right) - \bar{x}_{v_i} \quad (20)$$

$$M_i^{global} = \text{MultiHead}(\bar{x}_v, \bar{x}_v, \bar{x}_v), \quad (21)$$

where the subscript i denotes the current frame. The stride k controls the window size of the local neighbor frame. For both ends of the video frame sequence, we employ replicate padding strategy [49] to pad the original sequence length T_v to target sequence length $T_v + 2k$. The first frame is repeated k times for the beginning and the last frame is repeated k times for the end. For global difference information, we utilize multi-head attention [68] to capture the relative dependencies of all frames. The local-global embeddings $M = [M^{local}, M^{global}]$ passes through a Multi-Layer Perceptron (MLP) to predict the probability whether to mask the video frame. Formally,

$$\pi = \text{Softmax}(\text{MLP}(M)), \pi \in \mathbb{R}^{T_v \times 2}, \quad (22)$$

where the probability of index '0' ($\pi_{i,0}$) of π means to mask this video frame, and the probability of index '1' ($\pi_{i,1}$) means to keep this video frame. The subscript i represents i -th frame in the video clip. We can easily obtain the keyframe masking decision vector D by sampling from probability π and drop the uninformative frame $x_v = \bar{x}_v \odot D$ [62]. To ensure that the sparse video frame sequence \hat{x}_v and the original sequence x_v have similar semantics in the embedding space, we employ *gated recurrent units* GRU [19] and L2 regularization to compute video frame sequence reconstruction loss:

$$\mathcal{L}_{recon} = \| \text{GRU}(x_v) - \text{GRU}(\hat{x}_v) \|_2 \quad (23)$$

B EXPERIMENTS SETTING

B.1 Datasets

CMU-MOSI [81]. This dataset consists of 2199 videos, which contains manually transcribed text, audio and visual modal information. The training set, validation set, and test set each contained 1284, 229, and 686 samples. The label is a sentiment score (on a range of -3 to 3). Where sentiment score greater than 0 is positive, less than 0 is negative, and equal to 0 is neutral.

CMU-MOSEI [82]. The dataset collects of 22,856 videos from youtube, The dataset includes training dataset (16326 samples), the valid dataset (1871 samples) and the test dataset (4659 samples). The meaning of the label is the same as that of CMU-MOSI.

MELD [55]. It incorporates the same dialogues as EmotionLines, but introduces additional audio and visual modalities alongside text. Comprising over 1400 dialogues and 13000 utterances from the Friends TV series, MELD involves multiple speakers engaging in the dialogues. MELD provides sentiment annotations (positive, negative, and neutral) for each utterance. We utilize the multi-modal sentiment analysis datasets CMU-MOSI, CMU-MOSEI, and MELD to construct our training and testing sets. We train the model on the source domain and perform inference on both the source and target domains. We select to report the test set performance corresponding to the best performance observed on the validation set with 200 epochs. For datasets CMU-MOSI and CMU-MOSEI, we discretize the labels to obtain a three-class classification task. The distribution of labels (Negative, Neutral, Positive) in the three test sets are as follows : CMU-MOSI:{347, 106, 233}, MELD:{1015, 1891, 1685}, and CMU-MOSEI:{831, 1256, 521}. The three-class dataset exhibits approximate balance, and we report 3-class accuracy as the evaluation metric.

REFERENCES

- [1] Kartik Ahuja, Karthikeyan Shanmugam, Kush Varshney, and Amit Dhurandhar. 2020. Invariant risk minimization games. In *International Conference on Machine Learning*. PMLR, 145–155.
- [2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893* (2019).
- [3] Saeid Asgari, Aliasghar Khani, Fereshthe Khani, Ali Gholami, Linh Tran, Ali Mahdavi Amiri, and Ghassan Hamarneh. 2022. Masktune: Mitigating spurious correlations by forcing to explore. *Advances in Neural Information Processing Systems* 35 (2022), 23284–23296.
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. (2023).
- [5] Yogesh Balaji, Swami Sankaranarayanan, and Rama Chellappa. 2018. Metareg: Towards domain generalization using meta-regularization. *Advances in neural information processing systems* 31 (2018).
- [6] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of ChatGPT on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023* (2023).
- [7] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432* (2013).
- [8] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.
- [9] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 67–74.
- [10] Junbum Cha, Kyungjae Lee, Sungrae Park, and Sanghyuk Chun. 2022. Domain generalization by mutual-information regularization with pre-trained models. In *European Conference on Computer Vision*. Springer, 440–457.

- [11] Prithvijit Chattopadhyay, Yogesh Balaji, and Judy Hoffman. 2020. Learning to balance specificity and invariance for in and out of domain generalization. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX* 16. Springer, 301–318.
- [12] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. 2022. Compound domain generalization via meta-knowledge encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7119–7129.
- [13] Chaoqi Chen, Luyao Tang, Feng Liu, Gangming Zhao, Yue Huang, and Yizhou Yu. 2022. Mix and Reason: Reasoning over Semantic Topology with Data Mixing for Domain Generalization. *Advances in Neural Information Processing Systems* 35 (2022), 33302–33315.
- [14] Chaoqi Chen, Luyao Tang, Leitian Tao, Hong-Yu Zhou, Yue Huang, Xiaoguang Han, and Yizhou Yu. 2023. Activate and Reject: Towards Safe Domain Generalization under Category Shift. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11552–11563.
- [15] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18710–18719.
- [16] Liang Chen, Yong Zhang, Yibing Song, Ying Shan, and Lingqiao Liu. 2023. Improved Test-Time Adaptation for Domain Generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24172–24182.
- [17] Liang Chen, Yong Zhang, Yibing Song, Anton van den Hengel, and Lingqiao Liu. 2023. Domain generalization via rationale invariance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1751–1760.
- [18] Liang Chen, Yong Zhang, Yibing Song, Jue Wang, and Lingqiao Liu. 2022. Ost: Improving generalization of deepfake detection via one-shot test-time training. *Advances in Neural Information Processing Systems* 35 (2022), 24597–24610.
- [19] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
- [20] Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555* (2020).
- [21] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Hua Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instruct-clip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* 36 (2024).
- [22] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2978–2988.
- [23] Qi Dou, Daniel Coelho de Castro, Konstantinos Kamnitsas, and Ben Glocker. 2019. Domain generalization via model-agnostic learning of semantic features. *Advances in neural information processing systems* 32 (2019).
- [24] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. 2021. Improving multi-modal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059* (2021).
- [25] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. 2023. PMR: Prototypical Modal Rebalance for Multimodal Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20029–20038.
- [26] Tao Feng, Lizhen Qu, and Gholamreza Haffari. 2023. Less is More: Mitigate Spurious Correlations for Open-Domain Dialogue Response Generation Models by Causal Discovery. *Transactions of the Association for Computational Linguistics* 11 (2023), 511–530.
- [27] Muhammad Ghifary, David Balduzzi, W Bastiaan Kleijn, and Mengjie Zhang. 2016. Scatter component analysis: A unified framework for domain adaptation and domain generalization. *IEEE transactions on pattern analysis and machine intelligence* 39, 7 (2016), 1414–1430.
- [28] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis Philippe Morency, and Soujanya Poria. 2021. Bi-Bimodal Modality Fusion for Correlation-Controlled Multimodal Sentiment Analysis. In *ICMI 2021-Proceedings of the 2021 International Conference on Multimodal Interaction*. Association for Computing Machinery, Inc, 6–15.
- [29] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. 1122–1131.
- [30] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. 2008. Nonlinear causal discovery with additive noise models. *Advances in neural information processing systems* 21 (2008).
- [31] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. 2020. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*. PMLR, 292–302.
- [32] Shoubo Hu, Kun Zhang, Zhitang Chen, and Laiwan Chan. 2020. Domain generalization via multidomain discriminant analysis. In *Uncertainty in Artificial Intelligence*. PMLR, 292–302.
- [33] Ziniu Hu, Zhe Zhao, Xinyang Yi, Tiansheng Yao, Lichan Hong, Yizhou Sun, and Ed Chi. 2022. Improving multi-task generalization via regularizing spurious correlation. *Advances in Neural Information Processing Systems* 35 (2022), 11450–11466.
- [34] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. 2016. Binarized neural networks. *Advances in neural information processing systems* 29 (2016).
- [35] LIU Junjie, XU Zhe, SHI Runbin, Ray CC Cheung, and Hayden KH So. 2019. Dynamic Sparse Training: Find Efficient Sparse Network From Scratch With Trainable Masked Layers. In *International Conference on Learning Representations*.
- [36] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [37] Aditya Kusupati, Vivek Ramanujan, Raghav Somani, Mitchell Wortsman, Prateek Jain, Sham Kakade, and Ali Farhadi. 2020. Soft threshold weight reparameterization for learnable sparsity. In *International Conference on Machine Learning*. PMLR, 5544–5555.
- [38] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. 2018. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.
- [39] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. 2019. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1446–1455.
- [40] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [41] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. 2018. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European conference on computer vision (ECCV)*. 624–639.
- [42] Chen Liang, Simiao Zuo, Minshuo Chen, Haoming Jiang, Xiaodong Liu, Pengcheng He, Tuo Zhao, and Weizhu Chen. 2021. Super tickets in pre-trained language models: From model compression to improving generalization. *arXiv preprint arXiv:2105.12002* (2021).
- [43] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [44] Junjie Liu, Zhe Xu, Runbin Shi, Ray CC Cheung, and Hayden KH So. 2020. Dynamic sparse training: Find efficient sparse network from scratch with trainable masked layers. *arXiv preprint arXiv:2005.06870* (2020).
- [45] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2247–2256.
- [46] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. 2021. Progressive Modality Reinforcement for Human Multimodal Emotion Recognition From Unaligned Multimodal Sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2554–2562.
- [47] Sijie Mai, Ya Sun, and Haifeng Hu. 2022. Curriculum Learning Meets Weakly Supervised Modality Correlation Learning. *arXiv preprint arXiv:2212.07619* (2022).
- [48] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. 2023. Masked Motion Predictors are Strong 3D Action Representation Learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10181–10191.
- [49] Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, and Houqiang Li. 2023. Masked Motion Predictors are Strong 3D Action Representation Learners. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10181–10191.
- [50] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *International conference on machine learning*. PMLR, 10–18.
- [51] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78, 5 (2016), 947–1012.
- [52] Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. 2014. Causal discovery with continuous additive noise models. (2014).
- [53] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 1256–1272.
- [54] Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. 2021. Gradient starvation: A learning proclivity in neural networks. *Advances in Neural Information Processing Systems* 34 (2021), 1256–1272.
- [55] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508* (2018).

- [56] Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, and Tao Mei. 2023. Modality-agnostic debiasing for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24142–24151.
- [57] Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, and Tao Mei. 2023. Modality-agnostic debiasing for single domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24142–24151.
- [58] Francesco Quinzan, Ashkan Soleymani, Patrick Jaillet, Cristian R Rojas, and Stefan Bauer. 2023. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*. PMLR, 28468–28491.
- [59] Francesco Quinzan, Ashkan Soleymani, Patrick Jaillet, Cristian R Rojas, and Stefan Bauer. 2023. Drcfs: Doubly robust causal feature selection. In *International Conference on Machine Learning*. PMLR, 28468–28491.
- [60] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2020. NIH Public Access, 2359.
- [61] Alexandre Rame, Corentin Dancette, and Matthieu Cord. 2022. Fishr: Invariant gradient variances for out-of-distribution generalization. In *International Conference on Machine Learning*. PMLR, 18347–18377.
- [62] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. 2021. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems* 34 (2021), 13937–13949.
- [63] Mohammad Rastegari, Vicente Ordóñez, Joseph Redmon, and Ali Farhadi. 2016. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European conference on computer vision*. Springer, 525–542.
- [64] Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris Mooij. 2012. On causal and anticausal learning. *arXiv preprint arXiv:1206.6471* (2012).
- [65] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proc. IEEE* 109, 5 (2021), 612–634.
- [66] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.
- [67] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Learning Factorized Multimodal Representations. In *International Conference on Representation Learning*.
- [68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [69] Riccardo Volpi and Vittorio Murino. 2019. Addressing Model Vulnerability to Distributional Shifts Over Image Transformation Sets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [70] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. 2022. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering* (2022).
- [71] Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196* (2019).
- [72] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems* 33 (2020), 6256–6268.
- [73] Zhe Xu and Ray CC Cheung. [n. d.]. Accurate and Compact Convolutional Neural Networks with Trained Binarization. ([n. d.]).
- [74] DingKang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. 2022. Disentangled Representation Learning for Multimodal Emotion Recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1642–1651.
- [75] Fu-En Yang, Yuan-Chia Cheng, Zu-Yun Shiau, and Yu-Chiang Frank Wang. 2021. Adversarial teacher-student representation learning for domain generalization. *Advances in Neural Information Processing Systems* 34 (2021), 19448–19460.
- [76] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1009–1021.
- [77] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. 2023. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7617–7630.
- [78] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. 2023. MM-BigBench: Evaluating Multimodal Models on Multimodal Content Comprehension Tasks. *arXiv preprint arXiv:2310.09036* (2023).
- [79] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1103–1114.
- [80] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [81] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).
- [82] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.
- [83] Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. 2023. Learning Language-guided Adaptive Hyper-modality Representation for Multimodal Sentiment Analysis. *arXiv preprint arXiv:2310.05804* (2023).
- [84] Xianbing Zhao, Yixin Chen, Wanting Li, Lei Gao, and Buzhou Tang. 2022. MAG+: An Extended Multimodal Adaptation Gate for Multimodal Sentiment Analysis. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4753–4757.
- [85] Xianbing Zhao, Yixin Chen, Sicen Liu, Xuan Zang, Yang Xiang, and Buzhou Tang. 2023. TMMDA: A New Token Mixup Multimodal Data Augmentation for Multimodal Sentiment Analysis. In *Proceedings of the ACM Web Conference 2023*. 1714–1722.
- [86] Shuchang Zhou, Yuxin Wu, Zekun Ni, Xinyu Zhou, He Wen, and Yuheng Zou. 2016. Dorefa-net: Training low bandwidth convolutional neural networks with low bandwidth gradients. *arXiv preprint arXiv:1606.06160* (2016).
- [87] Yuwei Zhou, Xin Wang, Hong Chen, Xuguang Duan, and Wenwu Zhu. 2023. Intra-and Inter-Modal Curriculum for Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*. 3724–3735.