Facilitating Multimodal Classification via Dynamically Learning Modality Gap

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multimodal learning falls into the trap of the optimization dilemma due to the modality imbalance phenomenon, leading to unsatisfactory performance in real applications. A core reason for modality imbalance is that the models of each modality converge at different rates. Many attempts naturally focus on adjusting learning procedures adaptively. Essentially, the reason why models converge at different rates is because the difficulty of fitting category labels is inconsistent for each modality during learning. From the perspective of fitting labels, we find that appropriate positive intervention label fitting can correct this difference in learning ability. By exploiting the ability of contrastive learning to intervene in the learning of category label fitting, we propose a novel multimodal learning approach that dynamically integrates unsupervised contrastive learning and supervised multimodal learning to address the modality imbalance problem. We find that a simple yet heuristic integration strategy can significantly alleviate the modality imbalance phenomenon. Moreover, we design a learning-based integration strategy to integrate two losses dynamically, further improving the performance. Experiments on widely used datasets demonstrate the superiority of our method compared with state-of-the-art (SOTA) multimodal learning approaches. The code is available at `https://anonymous.4open.science/r/Dynamic_Modality_Gap_Learning`.

1 Introduction

Multimodal learning (MML) [3, 29, 35, 30, 11, 20, 24, 32, 12] integrates heterogeneous information from different modalities to build an effective way to perceive the world. Over the past decades, multimodal learning has made incredible progress [24, 11, 12] and become a hot research topic with a wide range of real applications including image caption [6], cross-modal retrieval [41], vision reasoning [25, 8], and so on.

In multimodal learning, several recent studies [29, 24] have revealed an interesting phenomenon, i.e., the performance of the multimodal model is far from the upper bound or even inferior to the uni-modal in certain situations. The root of this problem lies in the existence of the modality imbalance phenomenon [29]. Concretely, there commonly exists dominant modality and non-dominant modality in heterogeneous multimodal data. Multimodal learning usually adopts a uniform objective. Due to greediness [31], the optimization tends to dominant modality while neglecting the non-dominant one during joint training, thus leading to unsatisfactory performance in real applications.

Recently, many impressive works [29, 10, 24, 12, 18] have been proposed to address the modality imbalance problem. Early pioneering approaches, such as gradient blending (G-Blend) [29], on-the-fly gradient modulation (OGM) [24], adaptive gradient modulation (AGM) [18], and prototypical modality rebalance (PMR) [12], focus on designing customized learning strategies for different modalities to adjust the optimization of dominant and non-dominant modality. These methods demonstrate

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

that suppressing the optimization of the dominant modality can alleviate the modality imbalance problem to a certain extent. Besides, several attempts, including uni-modal teachers (UMT) [10] and balanced multimodal learning [31], try to introduce extra networks as an auxiliary module to facilitate multimodal learning.

Although the aforementioned approaches can boost performance in MML, these solutions are based on the phenomenon of inconsistent learning speed itself and do not study the underlying causes of modality imbalance. We can't help but ask what is the essential reason behind this phenomenon. Is there a bias in the process of fitting category labels for different modalities? We carry out a simple experiment on Kinetics-Sounds dataset to seek answers. We adopt two types of labels to explore the influence of fitting labels. The first type is one-hot labels which indicate the category of each sample, where the loss is denoted as L_S. The second type is label free, i.e., uniform label $1/c$ for all samples, where c denotes the number of categories. The second loss is defined as L_U. Furthermore, we define a mixed loss $0.7L_S+0.3L_U$ by combining one-hot labels and uniform labels. The accuracy is reported in Figure 1. From Figure 1, we can observe that with proper intervention by using

Figure 1: The influence of labels fitting on performance gaps (best view in color), where L_S and L_U denote the loss with one-hot labels and uniform labels (label free).

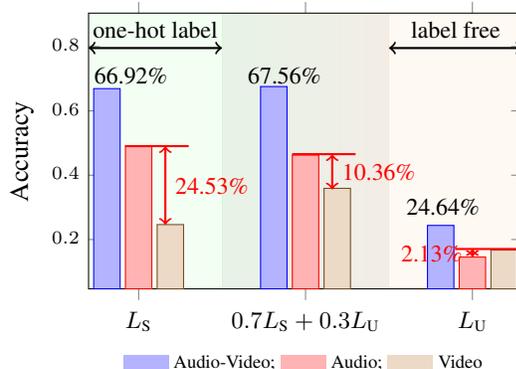
uniform labels, the performance is slightly better than the model that fits one-hot labels. More importantly, the performance gap becomes smaller if we learn from uniform labels. This means that the difference between audio and video modalities is smaller in feature space when we reduce the weight of one-hot labels. The experiment inspires us that appropriate intervention label fitting can alleviate the difference in the learning ability of different modalities. This also implies that fitting category labels is a core cause of modality imbalance in multimodal learning.

How do we impose positive intervention in multimodal learning so that the impact of fitting labels on modality imbalance is as low as possible without affecting the overall performance? For multimodal learning, although the models are learned from the heterogeneous data, we hope that multimodal data describing the same entity should be as close as possible in the feature space, which is usually modeled as contrastive learning [26]. Ideally, contrastive learning can also mitigate the effect of modality imbalance problem. Hence, we introduce contrastive learning to impose positive intervention in multimodal learning to alleviate the impact of fitting labels.

In this paper, we propose a novel multimodal learning approach by integrating unsupervised contrastive learning and supervised multimodal learning dynamically. Specifically, after demonstrating the effectiveness of unsupervised contrastive learning in multimodal learning, we design two dynamical integration strategies, i.e., a heuristic and a learning-based integration strategy. Our contributions are outlined as follows: (1). We observe a key phenomenon: fitting category labels leads to a larger performance gap between different modalities. To the best of our knowledge, this is the first time that the modality imbalance problem has been analyzed from the perspective of category label fitting. (2). We propose a novel multimodal learning approach by integrating unsupervised contrastive learning and supervised multimodal learning. Two strategies are designed for dynamic integration. (3). Extensive experiments on widely used datasets show that our proposed approach can significantly outperform other baselines to achieve state-of-the-art performance.

2   Related Work

2.1   Multimodal Learning

Multimodal learning aims to leverage multimodal data from different sources to improve model performance. Based on the fusion strategy, multimodal learning approaches can be categorized into early fusion [29, 34, 39, 38], late fusion [36, 35, 1, 21], and hybrid fusion [17, 40]. Early fusion methods

2

aim to integrate multimodal features to study the interrelationship between different modalities with joint representations when features are extracted by encoders. Representative early fusion methods include G-Blend [29], association-based fusion (AF) [20], and DOMFN [34]. On the contrary, late fusion methods leverage the prediction of each model to make final decisions. Late fusion methods can be divided into two categories, i.e., soft late fusion and hard late fusion, where the former utilizes the confidence score to make decisions and the latter the category decision of each model. Pioneering late fusion methods include modality-specific learning rate (MSLR) [35]. Hybrid fusion methods try to amalgamate the advantages of both early and late fusion methods. Representative hybrid fusion methods include multimodal transfer module (MMTM) [17] and balanced multi-modal learning [31]. Although these methods explore the algorithms and applications in multimodal learning, all of them assume that each modality can make sufficient contributions to achieve satisfactory performance during the training procedure.

2.2 Imbalanced Multimodal Learning

In reality, an obvious situation is that multimodal data and models are diverse, which naturally leads to different contributions during the training procedure. Recent works [29, 24, 12, 18] have shown that modality imbalance is a ubiquitous phenomenon and often results in unsatisfactory performance or even worse than unimodal algorithms in some cases. Considering the existence of dominant modality and non-dominant modality, early pioneering approaches [29, 24, 12] focus on adjusting learning speed for different modalities with customized learning strategies to balance the optimization of dominant and non-dominant modality. For instance, G-Blend [29] proposes to minimize the overfitting-to-generalization ratio (OGR) by using a gradient blending technique based on the modality's overfitting behavior. OGM [24] utilizes an on-the-fly gradient modulation strategy to control the modality's optimization procedure. To achieve the purpose of balanced multimodal learning, PMR [12] designs a prototypical modal rebalance strategy to facilitate the learning of non-dominant modality. Other attempts [10, 31] try to utilize extra networks to facilitate multimodal learning. Concretely, UMT [10] utilizes the teacher networks to distill the pretrained unimodal features to the multimodal network to tackle the modality imbalance problem. Balanced multimodal learning [31] utilizes the gradient norm and model parameters' norm to define conditional learning speed and uses it to guide the learning procedure. These methods alleviate the modality imbalance problem to a certain extent.

3 Methodology

We present our proposed method in this section. Specifically, we first present the problem definition of multimodal learning. Then, we introduce unsupervised contrastive learning to impose positive intervention in multimodal learning and propose two dynamical integration strategies to maximize the learning collaboration of unsupervised contrastive learning and supervised multimodal learning.

3.1 Preliminary

For the sake of simplicity, we use boldface lowercase letters like $\boldsymbol{a}$ and boldface uppercase letters like $\boldsymbol{A}$ to denote vectors and tensors, respectively. The i-th element of $\boldsymbol{a}$ is denoted as a_i. Furthermore, we use $\|\cdot\|_2$ to denote L_2 norm of the vectors.

The goal of multimodal learning is to train a model to predict the category labels for given multimodal data. Without any loss of generality, we use $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^n$ to denote the training data points, where each data point is with m modalities, i.e., $\boldsymbol{x}_i = \{\boldsymbol{x}_i^{(j)}\}_{j=1}^m$. The category labels are represented as $\boldsymbol{Y} = \{\boldsymbol{y}_i \mid \boldsymbol{y}_i \in \{0,1\}^c\}_{i=1}^n$, where c denotes the number of category labels.

For deep learning based multimodal approaches, we usually adopt a deep neural network to extract representation from original space into feature space. We utilize $\phi^{(j)}(\cdot)$ to denote the feature extraction function for j-th modality. Given data point $\boldsymbol{x}_i^{(j)}$, the feature extraction can be formed as:

$$\boldsymbol{z}_i^{(j)} = \phi^{(j)}(\boldsymbol{x}_i^{(j)}; \Phi^{(j)}),$$

where $\boldsymbol{z}_i^{(j)} \in \mathbb{R}^d$ denotes the d-dimension feature vector of $\boldsymbol{x}_i^{(j)}$, and $\Phi^{(j)}$ denotes the parameters of j-th encoder. After vectors for all modalities are extracted, we adopt a fusion function $f(\cdot)$ to fuse

3

the different feature vectors. Then, we leverage a fully-connected layer to map the vector into $\mathbb{R}^c$. This procedure can be formed as:

$$\boldsymbol{z}_i = f(\boldsymbol{z}_i^{(1)}, \cdots, \boldsymbol{z}_i^{(m)}), \quad \hat{\boldsymbol{y}}_i = \texttt{softmax}(\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}).$$

Here, $\boldsymbol{W} \in \mathbb{R}^{c \times D}$, $\boldsymbol{b} \in \mathbb{R}^c$ denote the weights and bias of the last fully-connected layer, respectively, and D denotes the dimension of $\boldsymbol{z}_i$. Then, the objective function of multimodal learning can be formulated as:

$$L_{\text{CLS}}(\boldsymbol{X}, \boldsymbol{Y}) = -\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{y}_i^{\top} \log \hat{\boldsymbol{y}}_i.$$

3.2 Integrating Unsupervised Contrastive Learning in MML

To bridge the heterogeneous data in feature space, we utilize contrastive learning [26] in multimodal learning. For a pair of data points $\{\boldsymbol{x}_i^{(j)}, \boldsymbol{x}_k^{(l)}\}$, we define the similarity as:

$$s(\boldsymbol{x}_i^{(j)}, \boldsymbol{x}_k^{(l)}) = \frac{[\boldsymbol{z}_i^{(j)}]^{\top} \boldsymbol{z}_k^{(l)}}{\|\boldsymbol{z}_i^{(j)}\|_2 \|\boldsymbol{z}_k^{(l)}\|_2}.$$

The modality matching objective function can be written as:

$$L_{\text{MM}}(\boldsymbol{X}) = -\frac{1}{2n_b} \sum_{i}^{n_b} \Big[\log \Big(\frac{\exp(s(\boldsymbol{x}_i^{(j)}, \boldsymbol{x}_i^{(l)})/\tau)}{\sum_k \exp(s(\boldsymbol{x}_i^{(j)}, \boldsymbol{x}_k^{(l)})/\tau)} \Big) + \log \Big(\frac{\exp(s(\boldsymbol{x}_i^{(j)}, \boldsymbol{x}_i^{(l)})/\tau)}{\sum_k \exp(s(\boldsymbol{x}_k^{(j)}, \boldsymbol{x}_i^{(l)})/\tau)} \Big) \Big],$$

where τ is the temperature parameter and n_b denotes the batch size. By integrating the classification loss and modality matching loss, we can get the following objective function:

$$L_{\text{Total}} = (1 - \alpha)L_{\text{CLS}}(\boldsymbol{X}, \boldsymbol{Y}) + \alpha L_{\text{MM}}(\boldsymbol{X}), \tag{1}$$

where α denotes the weighted parameter between two losses.

3.3 Dynamic Integration

Although a fixed value of α allows the model to take into account both classification loss and modality matching loss, it cannot dynamically evaluate the weight of two losses during training. Hence, we propose two strategies to adjust α dynamically to balance two losses.

Firstly, we utilize a monotonically decreasing function to adjust the impact of category labels. The definition of the function can be written as: $\alpha_t = \omega(t)$, where t denotes the number of training epochs. In this paper, we set $\omega(t) = 1 - e^{-\frac{1}{t}}$.

Then, we further exploit a learning-based integration method by utilizing bi-level optimization strategy [28]. Specifically, while considering optimizing the multimodal classification loss L_{CLS}, we use the minimum value of the total loss L_{Total} to restrict the feasible region of the parameters θ. In other words, we require the parameters not just to minimize classification loss but also to comply with a precisely defined constraint, i.e., simultaneously minimize a composite loss function—a strategically engineered combination of modality matching loss and multimodal classification loss. The specific formula is defined as follows:

$$\min_{0 \le \alpha \le 1} L_{\text{CLS}}(\theta^*(\alpha)) \quad \text{s.t.} \quad \theta^*(\alpha) \in \operatorname*{argmin}_{\theta} \big\{ (1 - \alpha)L_{\text{CLS}}(\theta) + \alpha L_{\text{MM}}(\theta) \big\}. \tag{2}$$

Here, θ denotes the parameters of multimodal models, and α emerges as a key parameter, delicately balancing modality matching loss and multimodal classification loss to direct the model toward an optimal balance where both types of loss are effectively managed. The optimal parameter set, $\theta^*(\alpha)$, thus represents a fine-tuned balance that, for any chosen α, strategically minimizes this composite loss. Within Equation (2), L_{CLS} is pivotal for guiding classification accuracy, while L_{MM} enhances the model's ability to establish meaningful connections across different modalities.

We utilize an approximation method proposed by [14] to solve bi-level optimization problem (2) efficiently. Specifically, the gradient of $L_{\text{CLS}}(\theta(\alpha))$ with respect to α can be approximated by:

$$\nabla L_{\text{CLS}}(\theta(\alpha)) = -\nabla_{\alpha,\theta}^2 L_{\text{Total}} [\nabla_{\theta,\theta}^2 L_{\text{Total}}]^{-1} \nabla_{\theta} L_{\text{CLS}}(\boldsymbol{X}, \boldsymbol{Y}). \tag{3}$$

4

Algorithm 1: The Proposed Algorithm.

Input : Training set $\mathcal{X}$, labels $\mathcal{Y}$, method.
Output : Learned parameters $\{\theta\}$ of all models.
INIT initialize parameters θ, parameter α, maximum iterations T, learning rate η_α.
for $t = 1$ **to** T **do**
    /* updating neural network parameters θ. */
    for $i = 1$ **to** *Inner_Iters* **do**
        Calculate total loss L_{Total} by forward phase.
        Update parameters θ according to its gradient.
    end
    /* updating weighting parameters α based on the chosen method. */
    if *method == 'learning-based'* **then**
        Calculate gradient apprtximation:
        $\nabla L_{\text{CLS}}(\theta(\alpha)) = -\nabla^2_{\alpha,\theta} L_{\text{Total}} \cdot [\nabla^2_{\theta,\theta} L_{\text{Total}}]^{-1} \cdot \nabla_\theta L_{\text{CLS}}(\boldsymbol{X}, \boldsymbol{Y})$.
        Update α according to: $\alpha = \alpha - \eta_\alpha \nabla L_{\text{CLS}}(\theta(\alpha))$.
        Clip α into $[0,1]$: $\alpha := \max(0, \min(1, \alpha))$.
    else if *method == 'heuristic'* **then**
        Update α according to: $\alpha = 1 - e^{-1/t}$.
    end
end

Based on the approximation Equation in (3), we can use the gradient descent method to optimize α.

After defining the updating strategy for α, we utilize an alternating algorithm between model parameters θ and α to perform model learning. Specifically, our algorithmic process iteratively refines the model parameters θ and the parameter α, employing a nested loop structure where the inner loop focuses on the currently given α to minimize total loss to update θ, and the outer loop updates α by function or bi-level policy to minimize classification losses. Through this structured optimization, the model achieves a delicate balance between multimodal matching and classification losses. The overall algorithm of our model is outlined in Algorithm (1), where we utilize $method$ to indicate the chosen updating strategy in practice.

4 Experiments

4.1 Experimental Setup

Datasets: We select five widely used datasets, including Kinetics-Sounds [2], CREMA-D [5], Sarcasm [4], Twitter2015 [37], and NVGesture [31] datasets, to validate our proposed method. Among these datasets, the Kinetics-Sounds and CREMA-D datasets consist of both audio and video modalities. The Kinetics-Sounds dataset, which contains 19,000 video clips categorized into 31 distinct actions, aims to advance video action recognition. It is divided into a training set of 15,000 clips, a validation set of 1,900 clips, and a test set of 1,900 clips. The CREMA-D dataset, encompassing 7,442 clips, is divided into six emotional categories to enhance speech emotion analysis, with 6,698 clips in the training set and 744 clips in the test set. Furthermore, the Sarcasm and Twitter2015 datasets consist of image and text modalities. The Sarcasm dataset offers a compilation of 24,635 text-image pairs, divided into 19,816 for the training set, 2,410 for the validation set, and 2,409 for the test set. The Twitter2015 dataset contains 5,338 text-image combinations from Twitter, with 3,179 in the training set, 1,122 in the validation set, and 1,037 in the test set. Lastly, the NVGesture dataset is used to construct research that goes beyond the limitation of two modalities. In this paper, we use RGB, Depth, and optical flow (OF) modalities for experiments, with 1,050 samples in the training set and 482 samples in the test set.

Baselines: We select a wide range of baselines for comparison. These baselines can be divided into two categories, i.e., traditional MML approaches and fusion methods with modal rebalancing strategies. The former category encompasses techniques like feature concatenation (CONCAT), affine transformation (Affine) [25], channel-wise fusion (Channel) [17], multi-layer LSTM fusion (ML-LSTM) [23], prediction summation (Sum), prediction weighting (Weight) [33], and enhanced trust modal combination (ETMC) [15]. And the latter category includes MSES [13], G-Blend [29], OGM [24], Greedy [30], DOMFN [34], MSLR [35], PMR [12], AGM [18], and MLA [39].

5

Table 1: Comparison with SOTA multimodal learning methods. The best results are highlighted in bold. The underlining symbol denotes the second best performance.

Method	Kinetics-Sounds		CREMA-D		Sarcasm		Twitter2015	
	ACC	MAP	ACC	MAP	ACC	F1	ACC	F1
Audio/Text	54.12%	56.69%	63.17%	68.61%	81.36%	80.65%	73.67%	68.49%
Video/Image	55.62%	58.37%	45.83%	58.79%	71.81%	70.73%	58.63%	43.33%
Concat	64.55%	71.31%	63.31%	68.41%	82.86%	82.43%	70.11%	63.86%
Affine	64.24%	69.31%	66.26%	71.93%	82.47%	81.88%	72.03%	59.92%
Channel	63.51%	68.66%	66.13%	71.75%	-	-	-	-
ML-LSTM	63.84%	69.02%	62.94%	64.73%	82.05%	70.73%	70.68%	65.64%
Sum	64.97%	71.03%	63.44%	69.08%	82.94%	82.47%	73.12%	66.61%
Weight	65.33%	71.33%	66.53%	73.26%	82.65%	82.19%	72.42%	65.16%
ETMC	65.67%	71.19%	65.86%	71.34%	83.69%	83.23%	73.96%	67.39%
MSES	64.71%	72.52%	61.56%	66.83%	84.18%	83.60%	71.84%	66.55%
G-Blend	67.12%	71.39%	64.65%	68.54%	83.35%	82.71%	74.35%	68.69%
OGM	66.06%	71.44%	66.94%	71.73%	83.23%	82.66%	<u>74.92%</u>	68.74%
Greedy	66.52%	72.81%	66.64%	72.64%	-	-	-	-
DOMFN	66.25%	72.44%	67.34%	73.72%	83.56%	82.62%	74.45%	68.57%
MSLR	65.91%	71.96%	65.46%	71.38%	84.23%	83.69%	72.52%	64.39%
PMR	66.56%	71.93%	66.59%	70.36%	83.61%	82.49%	74.25%	68.62%
AGM	66.02%	72.52%	67.07%	73.58%	<u>84.28%</u>	83.44%	74.83%	69.11%
MLA	<u>70.04%</u>	<u>74.13%</u>	<u>79.43%</u>	<u>85.72%</u>	84.26%	83.48%	73.52%	67.13%
Ours-H	69.05%	72.97%	72.15%	80.45%	84.12%	<u>83.98%</u>	73.87%	<u>69.17%</u>
Ours-LB	**72.53%**	**78.38%**	**83.62%**	**90.06%**	**84.97%**	**84.57%**	**75.01%**	**70.57%**

Evaluation Metrics: Following [24], we utilize accuracy (ACC) and mean Average Precision (MAP) as performance metrics for audio-video datasets. For text-image datasets, we adopt ACC and Macro F1-score (Mac-F1) [4, 37]. ACC measures the proportion of correct predictions to total predictions, indicating the overall predictive accuracy. Macro F1 calculates the average of F1 scores across all categories, balancing precision and recall to evaluate performance evenly across classes. MAP represents the average precision across all categories, assessing the model's ranking ability for each category.

Implementation Details: In our experiments, we utilize raw data for experiments. Following [24, 12], for the Kinetics-Sounds and CREMA-D datasets, ResNet18 [16] serves as the foundational architecture for processing both audio and video data. For video analysis, we select 10 frames from each clip and subsequently sample three frames uniformly as inputs. We adapt ResNet18's input channels from three to one to accommodate our data format [7]. In terms of audio, we transform our sound recordings into spectrograms measuring 257×1004 for Kinetics-Sounds and 257×299 for CREMA-D, employing the librosa [22] library for conversion. For text-image datasets, our framework incorporates ResNet50 for images and BERT [9] for text processing. We resize images to 224×224 and limit text sequences to a maximum length of 128 characters. Optimization for the audio-video datasets is conducted using stochastic gradient descent (SGD) with a momentum set to 0.9 and a weight decay parameter of 10^{-1}. We initialize the learning rate to 10^{-2}, progressively reducing it by a factor of ten upon observing a plateau in loss reduction, with a batch size of 256. For text-image datasets [4, 37], we employ the Adam optimizer starting with a learning rate of 10^{-4}, with a batch size of 128. All models are trained on a single RTX 3090 GPU.

4.2 Comparison with SOTA MML Baselines

Main results on audio-video datasets and image-text datasets are presented in Table 1, where "Our-H" and "Ours-LB" denote the proposed method based on heuristic strategy and learning-based strategy, respectively. From the results, we can derive the following observation: (1). Compared with all baselines including traditional multimodal learning approaches and fusion methods with modal rebalancing strategies, our proposed method with learning-based strategy can achieve best performance by a large margin on all datasets. We can also find that the model with learning-based strategy can achieve better performance than that with heuristic strategy. (2). Across the Twitter2015 dataset, there is a discernible trend where the optimal unimodal performance outstrips that of multimodal joint

6

Table 2: Results on NVGesture dataset.

Method	ACC	F1
MSES	81.12%	81.47%
OGR-GB	82.99%	83.05%
MSLR	82.37%	82.39%
AGM	82.78%	82.84%
MLA	<u>83.73</u> %	83.87%
Ours-H	83.24%	<u>83.87%</u>
Ours-LB	**84.36%**	**84.68%**

Table 3: Ablation study on Kinetics-Sounds dataset. The symbols "CL" and "DI" denote that whether the contrastive learning and dynamic integration are applied during training.

Module		MAP			
CL	DI	Audio-Video	Audio	Video	GAP
×	×	69.32%	48.82%	27.19%	21.63%
✓	×	71.76%	51.05%	47.05%	3.80%
✓	✓	**78.97%**	**58.40%**	**60.42%**	**2.02%**

Table 4: Results with different dynamic integration strategy on Kinetics-Sounds dataset.

Modal	Constant			Stepwise				Dynamic	
	0	0.5	1	$h(0)$	$h(1)$	$h(0.05)$	$h(0.95)$	Ours-H	Ours-LB
Audio-Video	64.55%	64.70%	28.67%	65.17%	66.92%	66.01%	67.41%	69.32%	72.89%
Audio	49.17%	46.30%	34.11%	51.12%	52.34%	52.21%	53.41%	53.89%	54.32%
Video	24.64%	44.02%	28.41%	41.21%	41.45%	42.31%	46.72%	49.18%	54.17%

learning. Additionally, in other datasets, fusion methodologies devoid of rebalancing mechanisms manifest negligible enhancements relative to the foremost unimodal performance, notably on the CREMA-D and Sarcasm datasets. This shortfall originates from the prevalent challenge of modal imbalance. (3). Every modality rebalancing technique demonstrates significant improvements over traditional feature concatenation fusion. This finding not only underscores the detrimental impact of modal imbalance on performance but also corroborates the efficacy of the modality rebalancing approach. Detailed results with error bars are provided in the supplementary materials due to space limitations.

In Table 2, we report the comparison with SOTA baselines on NVGesture dataset, where the similar symbols are used to denote our method. From Table 2, we can see that differing from modal rebalancing methods restricted to scenarios with only two modalities, such as Greedy, our approach with learning-based strategy can address challenges in scenarios involving more than two modalities and achieve best results. Furthermore, our proposed method with heuristic strategy can also outperform all baselines in all cases.

4.3 Ablation Study

To comprehensively assess the effectiveness of our proposed method, we conduct experiments to study the influence of main components, i.e., contrastive learning (CL) and dynamic integration (DI). The results are shown in Table 3, where the "CL" and "DI" denote that whether the contrastive learning and dynamic integration are applied during training. Please note that dynamic integration depends on the contrastive learning loss. From Table 3, we can see that both contrastive learning and dynamic integration can boost performance in multimodal learning. Moreover, by integrating contrastive learning into multimodal learning, the performance gap between audio and video is greatly reduced. More results are presented in the supplementary materials due to space limitations.

4.4 Effectiveness of Integration Learning

Analysis of Integration Strategy: We further study the impact of different integration strategies for contrastive loss and classification loss. Specifically, we analyze three categories of integration strategy, i.e., constant, stepwise and dynamic strategy. For constant strategy, we assign a constant value to α for the experiment and run three sets of experiments by setting $\alpha = 0$, $\alpha = 0.5$, and $\alpha = 1$. Here, "$\alpha = 0$" denotes that we only perform supervised multimodal learning. Similarly, "$\alpha = 1$" denotes that we only perform unsupervised contrastive learning. For stepwise strategy, we define an indicator function $h(p)$, where $h(p)$ denotes that $\alpha = p$ if the current epoch is less than half

7

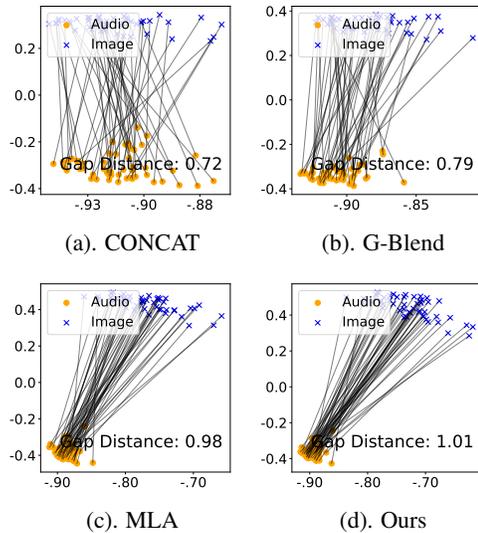



Figure 2: Visualizations of the modality gap distance on the CREMA-D dataset.

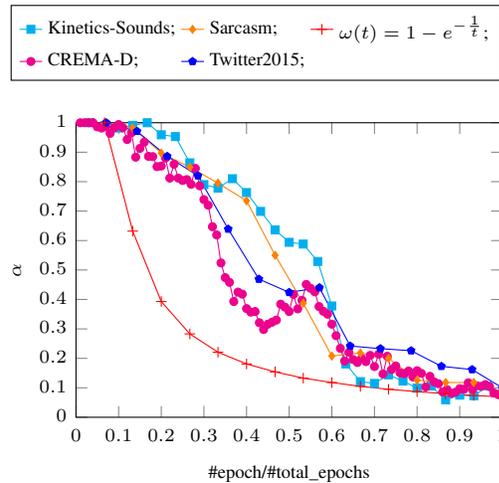



Figure 3: Change of α on different datasets. We illustrate the value of the heuristic integration strategy for comparison.

of the total epochs, otherwise $\alpha = 1 - p$. For dynamic strategy, we assign value to α by heuristic strategy (Ours-H) and learning-based strategy (Ours-LB).

The experimental results are shown in Table 4. From Table 4, we can draw the following observations: (1). Integrating multimodal learning and contrastive learning simultaneously with a constant ratio can slightly boost performance in some cases and greatly reduce the performance gap between audio and video. (2). In general, the model with a stepwise strategy can outperform the model with a constant strategy. Furthermore, we can find that the performance of the model with two-stage training, i.e., $h(0)$ or $h(1)$, is worse than that of the model which combines two losses with a constant value, i.e., $h(0.05)$ or $h(0.95)$. (3). The overall performance of the model with the dynamic strategy is better than that with the other strategy. Moreover, the model with the learning-based strategy can achieve the best performance. And the performance gap of this model is nil or negligible. The experimental results prove that the smaller the performance gap of the uni-modals, the better the overall performance of the model. More experimental results on CREMA-D dataset are provided in the supplementary materials due to space limitation.

Change of α for Learning-based Strategy: To further observe the change of the optimal α during training, we illustrate the change of α on all datasets. The results are shown in Figure 3, where α is calculated by learning-based strategy, and $\omega(t)$ denotes the heuristic based strategy. As the total epochs for different datasets are different, we change the x-axis as the proportion of the current epoch, i.e., #epoch/#total_epochs. From Figure 3, we can draw the following observations: (1). The general trend of the change for α is roughly the same on different datasets. (2). The customized function $\omega(t)$ is close to the actual changes to some extent, but there is still a gap between the customized function and the actual changes. In practice, it is difficult to fit the change of parameter α perfectly. Hence, we can see that our method has good adaptability in different scenarios.

4.5 Further Analysis

Analysis of Modality Gap: As mentioned in the paper [19], modality gap characterizes the correlation between different modalities in multimodal learning. And large modality gap leads to better performance in some situations. We further illustrate the modality gap for CONCAT, G-Blend, MLA, and our method. The results are shown in Figure 2. From Figure 2, we can find that our method can learn more discriminative representations and results in higher accuracy with a large modality gap compared with other methods.

Robustness Analysis of the Pretrained Mode: We further exploit the robustness of the CLIP [26] model on Sarcasm and Twitter2015 datasets. We replace the encoders for image and text as the corre-

8

Table 5: Results on the Sarcasm and Twitter2015 datasets achieved by using the CLIP pre-trained model as encoders.

Method	Sarcasm			Twitter2015		
	Image	Text	Multi	Image	Text	Multi
CLIP	74.82%	82.15%	83.11%	54.48%	71.75%	72.52%
CLIP+MLA	77.45%	83.19%	84.45%	56.53%	72.37%	73.95%
CLIP+Ours	**79.78%**	**83.67%**	**85.42%**	**64.67%**	**72.59%**	**74.43%**

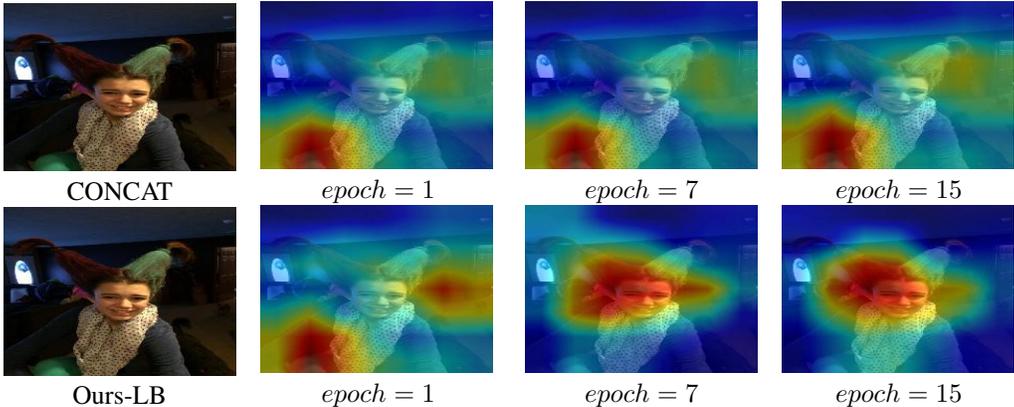



Figure 4: Visualization on Twitter2015 dataset. Our proposed method tends to perform feature learning first and then fit the learned features to the category labels.

sponding encoders pretrained by CLIP and fine-tune the model on Sarcasm and Twitter2015 datasets respectively. The results are shown in Table 5, where "CLIP+MLA" and "CLIP+Ours" present that we apply the MLA's and ours algorithm, respectively. From Table 5, we can draw the following observations: (1). Both CLIP+MLA and CLIP+Ours can outperform CLIP in all cases. (2). With the help of dynamic integration, the performance of our method is better than that of MLA.

Visualization: We utilize GradCAM [27] to showcase the visualization of image regions that attract the weak modality's focus during training. By using GradCAM, the importance scores are assigned to every pixel in each feature map, aiding in identifying the image regions critical for the model's predictions. We compare the visualization for CONCAT and our proposed method. The visualization results are presented in Figure 4, where the second, the third, and the last columns denote the results of the first, the seventh, and the last epoch, respectively. The category label for this image is "Negative" and the corresponding text is "Crazy hair day ! T is a contender.". By comparing our method with CONCAT, we can see that our method focuses on the textual information from text modality, and then fits the learned features to the category labels.

5 Conclusion

In this paper, we discuss a core reason for modality imbalance in multimodal learning, i.e., fitting category labels. We find that appropriate positive intervention label fitting can correct the difference in learning ability for different modalities, thus alleviating the modality imbalance phenomenon. Based on this observation, we propose a novel multimodal learning approach to overcome modality imbalance problem by dynamically integrating unsupervised contrastive learning and supervised multimodal learning. We design a heuristic strategy and a learning based strategy to perform integration dynamically. Experiments on various datasets demonstrate that our method can boost performance in multimodal learning.

For the **limitations** of our proposed method. The root cause of modality imbalance caused by fitting category labels is worth discussing in depth. Does the specific category label contain attributes that are more suitable for fitting a certain modality? We leave it as a future work.

9

References

[1] M. Alfaro-Contreras, J. J. Valero-Mas, J. M. Iñesta, and J. Calvo-Zaragoza. Late multimodal fusion for image and audio music transcription. *ESWA*, 216:119491, 2023.

[2] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *ICCV*, pages 609–617. IEEE, 2017.

[3] T. Baltrusaitis, C. Ahuja, and L. Morency. Multimodal machine learning: A survey and taxonomy. *TPAMI*, 41(2):423–443, 2019.

[4] Y. Cai, H. Cai, and X. Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pages 2506–2515. Association for Computational Linguistics, 2019.

[5] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma. CREMA-D: crowd-sourced emotional multimodal actors dataset. *TAC*, 5(4):377–390, 2014.

[6] A. X. Chang, W. Monroe, M. Savva, C. Potts, and C. D. Manning. Text to 3d scene generation with rich lexical grounding. In *ACL*, pages 53–62. The Association for Computer Linguistics, 2015.

[7] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE, 2020.

[8] L. Chen, B. Li, S. Shen, J. Yang, C. Li, K. Keutzer, T. Darrell, and Z. Liu. Large language models are visual reasoning coordinators. In *NeurIPS*, 2023.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186. Association for Computational Linguistics, 2019.

[10] C. Du, T. Li, Y. Liu, Z. Wen, T. Hua, Y. Wang, and H. Zhao. Improving multi-modal learning with uni-modal teachers. *CoRR*, abs/2106.11059, 2021.

[11] C. Du, J. Teng, T. Li, Y. Liu, Y. Wang, Y. Yuan, and H. Zhao. Modality laziness: Everybody's business is nobody's business, 2022.

[12] Y. Fan, W. Xu, H. Wang, J. Wang, and S. Guo. PMR: prototypical modal rebalance for multimodal learning. In *CVPR*, pages 20029–20038. Computer Vision Foundation / IEEE, 2023.

[13] N. Fujimori, R. Endo, Y. Kawai, and T. Mochizuki. Modality-specific learning rate control for multimodal classification. In *ACPR*, volume 12047, pages 412–422, 2019.

[14] S. Ghadimi and M. Wang. Approximation methods for bilevel programming. *CoRR*, abs/1802.02246, 2018.

[15] Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification with dynamic evidential fusion. *TPAMI*, 45(2):2551–2566, 2023.

[16] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. Computer Vision Foundation / IEEE, 2016.

[17] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida. MMTM: multimodal transfer module for CNN fusion. In *CVPR*, pages 13286–13296. Computer Vision Foundation / IEEE, 2020.

[18] H. Li, X. Li, P. Hu, Y. Lei, C. Li, and Y. Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pages 22157–22167. IEEE, 2023.

[19] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

[20] X. Liang, Y. Qian, Q. Guo, H. Cheng, and J. Liang. AF: an association-based fusion method for multi-modal classification. *TPAMI*, 44(12):9236–9254, 2022.

10

[21] B. Liu, L. He, Y. Xie, Y. Xiang, L. Zhu, and W. Ding. Minjot: Multimodal infusion joint training for noise learning in text and multimodal classification problems. *INFFUS*, 102:102071, 2024.

[22] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto. librosa: Audio and music signal analysis in python. In *Python in Science Conference*, pages 18–24. scipy.org, 2015.

[23] W. Nie, Y. Yan, D. Song, and K. Wang. Multi-modal feature fusion based on multi-layers LSTM for video emotion recognition. *MTA*, 80(11):16205–16214, 2021.

[24] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pages 8228–8237. Computer Vision Foundation / IEEE, 2022.

[25] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, pages 3942–3951. AAAI Press, 2018.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139, pages 8748–8763. PMLR, 2021.

[27] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *IJCV*, 128(2):336–359, 2020.

[28] L. N. Vicente and P. H. Calamai. Bilevel and multilevel programming: A bibliography review. *JGO*, 5(3):291–306, 1994.

[29] W. Wang, D. Tran, and M. Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12692–12702. Computer Vision Foundation / IEEE, 2020.

[30] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pages 24043–24055. PMLR, 2022.

[31] N. Wu, S. Jastrzebski, K. Cho, and K. J. Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, volume 162, pages 24043–24055. PMLR, 2022.

[32] P. Xu, X. Zhu, and D. A. Clifton. Multimodal learning with transformers: A survey. *TPAMI*, 45(10):12113–12132, 2023.

[33] Y. Yang, K. Wang, D. Zhan, H. Xiong, and Y. Jiang. Comprehensive semi-supervised multi-modal learning. In *IJCAI*, pages 4092–4098. ijcai.org, 2019.

[34] Y. Yang, J. Zhang, F. Gao, X. Gao, and H. Zhu. DOMFN: A divergence-orientated multi-modal fusion network for resume assessment. In *ACMMM*, pages 1612–1620. ACM, 2022.

[35] Y. Yao and R. Mihalcea. Modality-specific learning rates for effective multimodal additive late-fusion. In *ACL*, pages 1824–1834. Association for Computational Linguistics, 2022.

[36] Y. Ye, X. Liu, Q. Liu, X. Guo, and J. Yin. Incomplete multiview clustering via late fusion. *CIN*, 2018:6148456:1–6148456:11, 2018.

[37] J. Yu and J. Jiang. Adapting BERT for target-oriented multimodal sentiment classification. In *IJCAI*, pages 5408–5414. ijcai.org, 2019.

[38] Y. Zeng, W. Yan, S. Mai, and H. Hu. Disentanglement translation network for multimodal sentiment analysis. *INFFUS*, 102:102031, 2024.

[39] X. Zhang, J. Yoon, M. Bansal, and H. Yao. Multimodal representation learning by alternating unimodal adaptation. *CoRR*, abs/2311.10707, 2023.

[40] X. Zheng, C. Tang, Z. Wan, C. Hu, and W. Zhang. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, pages 11381–11389. AAAI Press, 2023.

[41] L. Zhu, T. Wang, F. Li, J. Li, Z. Zhang, and H. T. Shen. Cross-modal retrieval: A systematic review of methods and future directions. *CoRR*, abs/2308.14263, 2023.

11

416
417

Supplementary Materials for Facilitating Multimodal Classification via Dynamically Learning Modality Gap

418

A Additional Experimental Results

419

A.1 Comparison with SOTA MML Baselines

420

To remove the randomness, we run the experiment 3 times with different random seeds and present the detailed performance with $mean$ and $std.$ values in Table A1.

Table A1: Detailed performance with $mean$ and $std.$ values.

Dataset	Ours-H		Ours-LB	
	ACC	MAP	ACC	MAP
Kinetics-Sounds	69.05%±0.15%	72.97%±0.43%	72.53%±0.31%	78.38%±0.37%
CREMA-D	72.15%±0.32%	80.45%±0.85%	83.62%±0.11%	90.06%±1.09%
	ACC	F1	ACC	F1
Sarcasm	84.12%±0.17%	83.98%±0.22%	84.97%±0.27%	84.57%±0.18%
Twitter2015	73.87%±0.35%	69.17%±0.26%	75.02%±0.16%	70.57%±0.28%
NVGesture	83.24%±0.07%	83.87%±0.18%	84.36%±0.14%	84.68%±0.24%

421

422

A.2 Ablation Study

423
424
425

In Table A2, we report the accuracy on all datasets except Kinetics-Sounds dataset for the ablation study. From Table A2, we can find that on CREMA-D, Sarcasm, and Twitter2015 datasets, contrastive learning and dynamic integration can boost performance in multimodal learning. Moreover, the performance gap is reduced by integrating contrastive learning into multimodal learning.

Table A2: Ablation study on the rest datasets. The "CL" and "DI" denote that whether contrastive learning and dynamic integration are applied during training.

Dataset	Module		MAP			
	CL	DI	Multi	Audio	Video	GAP
	×	×	76.07%	70.97%	34.15%	36.82%
CREMA-D	✓	×	86.32%	72.11%	52.51%	19.06%
	✓	✓	**90.06%**	**75.27%**	**67.36%**	**7.91%**

Dataset	Module		F1			
	CL	DI	Multi	Image	Text	GAP
	×	×	82.43%	62.81%	77.96%	15.15%
Sarcasm	✓	×	83.10%	68.74%	80.72%	11.98%
	✓	✓	**84.57%**	**74.53%**	**83.03%**	**8.50%**
	×	×	63.86%	40.99%	68.38%	27.39%
Twitter2015	✓	×	65.33%	46.84%	69.04%	22.20%
	✓	✓	**70.57%**	**53.43%**	**69.68%**	**16.25%**

426

427

A.3 Analysis of Integration Strategy on CREMA-D dataset.

428
429
430
431

We further report the impact of different integration strategy on CREMA-D dataset. The results are shown in Table A3. On CREMA-D dataset, we can still observe similar results. Roughly speaking, the model with dynamic integration can achieve better performance compared with other strategies. Furthermore, the model with learning based strategy can achieve the best performance in all cases.

12

Table A3: Results with different dynamic integration strategy on CREMA-D dataset.

Modal	Constant			Stepwise				Dynamic	
	0	0.5	1	$h(0)$	$h(1)$	$h(0.05)$	$h(0.95)$	Ours-H	Ours-LB
Multi	63.31%	70.45%	26.49%	66.45%	70.24%	69.11%	71.45%	72.39%	84.11%
Video	18.68%	42.54%	20.42%	45.14%	49.97%	46.41%	55.32%	57.14%	64.89%
Audio	55.65%	60.17%	33.15%	56.19%	57.38%	58.09%	60.18%	61.89%	65.13%

13

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Main contributions and scope were reflected in the abstract and introduction (in the last paragraph of introduction, page 2).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations of the work was discussed in section 5.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper doesn't involve theoretical result.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental setup, including dataset, baseline, evaluation metric and implementation details, was provided in section 4.1. In the implementation details part, we provide the network architecture, input data processing, optimization algorithm, the value of main parameters, hardware environment, and so on, to support the reproducing of the method.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release our code at `https://anonymous.4open.science/r/Dynamic_Modality_Gap_Learning`. Furthermore, all datasets we used in this paper are available online based on their corresponding paper.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: All implementation details were provided in section 4.1.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We run the main experiments 3 times with different random seeds. The results with error bars were reported in section A.1 in the supplementary materials.

8. **Experiments Compute Resources**

14

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: The computer resource information for all experiments was provided in the section 4.1.

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We have checked the NeurIPS code of ethics for our paper.

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: Our paper does not deal with this aspect of the problem.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our paper does not deal with this aspect of the problem.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The assets (datasets, code, models) used in the paper are open source, and our use follows the relevant protocols.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The code we released in this paper contains related documents.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper doesn't involve the crowdsourcing experiments and research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper doesn't involve the potential risks.

15