Balancing Multimodal Learning via Online Logit Modulation

Daoming Zong, **Chaoyue Ding**, **Baoxiang Li**[*], **Jiakui Li** and **Ken Zheng**
SenseTime Research
{ecnuzdm, cydingcs}@gmail.com, {libaoxiang, lijiakui, zhengken}@sensetime.com

Abstract

Multimodal learning is *provably* superior to unimodal learning. However, in practice, the best-performing unimodal networks often outperform jointly trained multimodal networks. This phenomenon can be attributed to the varying convergence and generalization rates across different modalities, leading to the dominance of one modality and causing underfitting of other modalities in simple multimodal joint training. To mitigate this issue, we propose two key ingredients: *i*) disentangling the learning of unimodal features and multimodal interaction through an intermediate representation fusion block; *ii*) modulating the logits of different modalities via dynamic coefficients during training to align their magnitudes with the target values, referred to as *online logit modulation* (`OLM`). Remarkably, `OLM` is model-agnostic and can be seamlessly integrated with most existing multimodal training frameworks. Empirical evidence shows that our approach brings significant enhancements over baselines on a wide range of multimodal tasks, covering video, audio, text, image, and depth modalities.

1 Introduction

Intuitively, multimodal models that fuse different modality data are expected to outperform unimodal models due to the richer information they provide. However, a counterintuitive finding is often observed in practice, where the best-performing unimodal networks outperform jointly trained multimodal networks, especially in coarse-grained multimodal classification tasks [Wang *et al.*, 2020; Peng *et al.*, 2022]. This phenomenon can be attributed to the fact that *different modalities learn representations at different rates of convergence and generalization*. Current mainstream multimodal joint training frameworks, which incorporate *late fusion* [Xu *et al.*, 2023a] to encode different modalities' features into a shared latent space and map them to the task space, may lead to inconsistent final convergence states among modalities, where some may be overfitted while others remain underfitted. In such cases, the multimodal model experiences degradation, relying excessively on a specific modality and failing to effectively utilize information from other modalities.

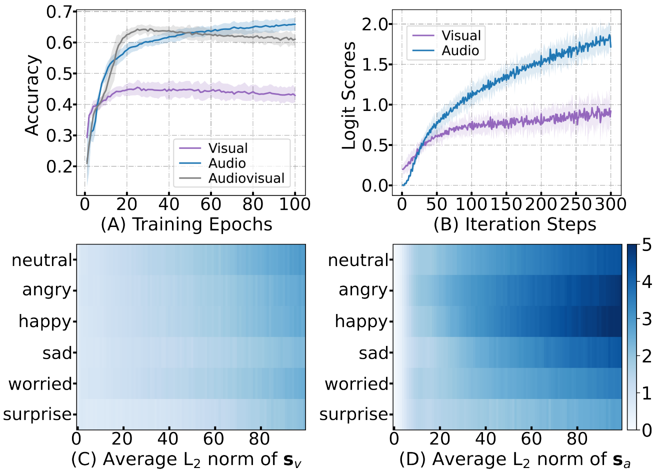


Figure 1: (A) Audiovisual and unimodal test accuracy on MER-MULTI. (B) The batch-average unimodal logit scores. (C-D) The averaged L_2 norms of unimodal logit vectors.

Existing research on balancing multimodal learning can be categorized into two approaches: *i*) *overcoming architectural deficiencies* and *ii*) *harmonizing multimodal training schemes*. The former involves designing modality-specific encoders to learn the unique features of each modality and modality-invariant encoders to learn shared representations across modalities [Hazarika *et al.*, 2020], or bridging the gap between modalities by employing modality-aware encoders [Xiao *et al.*, 2020] (*e.g.*, SlowFast [Feichtenhofer *et al.*, 2019], comprising a slow pathway operating at a low frame rate to capture spatial semantics and a Fast pathway operating at a high frame rate to capture fine temporal motion). The latter includes methods such as **G-blend** [Wang *et al.*, 2020], which computes an optimal blending of modalities based on their overfitting behaviors; **CUR** [Wu *et al.*, 2022], which balances the conditional learning rates between modalities using the condition utilization ratio to measure the performance difference between unimodal and bimodal models;

[*] Corresponding author

and **OGM-GE** [Peng *et al.*, 2022], which achieves dynamic and differentiated parameter updates through on-the-fly gradient modulation. Despite the modest success of these methods, they either require additional complex gradient computations and modifications to standard backpropagation or involve multiple training runs to estimate imbalanced proxies.

We postulate that joint multimodal training can only benefit from multimodal interactions when sufficient learning of unimodal features is ensured. To achieve this, we first decouple the learning of unimodal features and cross-modal feature interactions through an **I**ntermediate **R**epresentation **F**usion **B**lock (IRFB). Next, to coordinate the optimization processes of different modalities, we propose **O**nline **L**ogit **M**odulation (OLM). Our motivation arises from observations in Fig. 1 (a) and (b), where the audio modality contributes significantly to overall performance. This dominance of the audio modality is evident from the average unimodal logit scores within each batch, while the visual modality persistently remains under-optimized throughout training. Further observations in Fig. 1 (c) and (d) demonstrate that the logit norms corresponding to each category continuously increase during training, and the logit norms of the audio and visual modalities gradually diverge as training proceeds. Guided by these observations, we propose to use OLM to regulate the magnitudes of logit vectors across diverse modalities. The core principle of OLM involves the adaptive adjustment of logit norms for each modality during training. This is achieved by applying adaptive *logit coefficients* to the modality-wise logit vectors. The intention is to align the magnitude of the logit vector with its predetermined target. Consequently, this approach attenuates logit norms of swiftly converging modalities while amplifying those of modalities that converge at a slower pace. The cumulative effect engenders a more harmonized optimization process. To sum up, the contributions of this work are as follows:

- We decouple the learning of unimodal features and multimodal interactions via an intermediate representation fusion block (IRFB), thereby enabling sufficient training of unimodal features.
- We introduce OLM, an online logit calibration strategy that scales each modality logit vector during training, thus aligning their magnitudes with the target logit norms. Notably, OLM is model-agnostic and harmonizes seamlessly with most multimodal training architectures.
- Empirical findings substantiate that our approach significantly enhances the performance of baseline models across diverse multimodal tasks, including human action recognition, scene categorization, audiovisual event localization, and multimodal sentiment analysis.

2 Related Work

Multimodal Alignment and Fusion Features from distinct modalities typically inhabit separate embedding spaces [Li *et al.*, 2021; 2022]. The primary goal of multimodal alignment is to project these diverse modalities onto a shared representation space, thereby facilitating the modeling of subsequent cross-modal fusion/interaction. Contrastive learning [Chen *et al.*, 2020b] has extensively been employed to train transformer-based multimodal models for achieving modality alignment [Jia *et al.*, 2021; Radford *et al.*, 2021; Yang *et al.*, 2021; Li *et al.*, 2022; Shen *et al.*, 2023]. Moreover, beyond contrastive learning, the moment-based maximum mean discrepancy [Gretton *et al.*, 2012], and the optimal transport dataset distance [Alvarez-Melis and Fusi, 2020] have also been explored for cross-modal distributional alignment. Multimodal interaction can occur at three levels: input (a.k.a, *early fusion*), intermediate representation (a.k.a, *middle fusion*), and prediction or decision level (a.k.a, *late fusion*) [Xu *et al.*, 2023a]. Early fusion immediately combines features right after their extraction, often achieved through concatenation or summation of diverse representations. In contrast, in a late fusion setting, all modalities are trained independently and merged right before the model makes a decision. Middle fusion, which typically employs cross-modal attention or co-attention [Lu *et al.*, 2019] and its variants, such as 'attention bottleneck' [Nagrani *et al.*, 2021], enable more fine-grained modal interactions and yield more robust multimodal contextual representations.

Imbalanced Multimodal Learning Simple multimodal joint training can lead to the optimization of only one dominant modality, while other modalities suffer from underfitting [Wang *et al.*, 2020; Peng *et al.*, 2022; Huang *et al.*, 2022; Wu *et al.*, 2022; Fan *et al.*, 2023]. In such cases, the multimodal model excessively relies on a single modality, compromising its generalization performance. To mitigate such imbalanced optimization, a series of multimodal calibration training algorithms have been proposed. For instance, Wang *et al.* present **G-Blend**, which utilizes five-fold cross-validation to estimate the overfitting-to-generalization ratio and re-weights the training losses accordingly. Despite its effectiveness, this approach requires additional data splitting and training of individual unimodal models, resulting in an increased computational burden. Peng *et al.* propose on-the-fly gradient modulation (**OGM-GE**), which adaptively controls the optimization for each modality by monitoring their contributions to the learning objective. Moreover, it incorporates additional Gaussian noise to mitigate potential generalization degradation resulting from gradient modulation. Nevertheless, this method requires sampling from the distribution of gradient variances, hindering its training efficiency. Wu *et al.* introduce the concept of the conditional utilization rate (**CUR**), which is defined as the accuracy gain when merging one modality with another into a model and is applied to update one of the unimodal branches intentionally. However, it complicates the training protocol due to its iterative loops between standard training steps and rebalancing steps. **PMR** [Fan *et al.*, 2023] leverages the prototypes, namely the centroids of each modal in representation space, to adjust the learning direction of each modal towards its prototypes. Different from these methods, OLM accelerates the slow-learning modality and alleviates the suppression from the dominant modality by adaptively modulating the logit magnitude (while maintaining the learning directions unchanged) of each modality during the entire training stage.

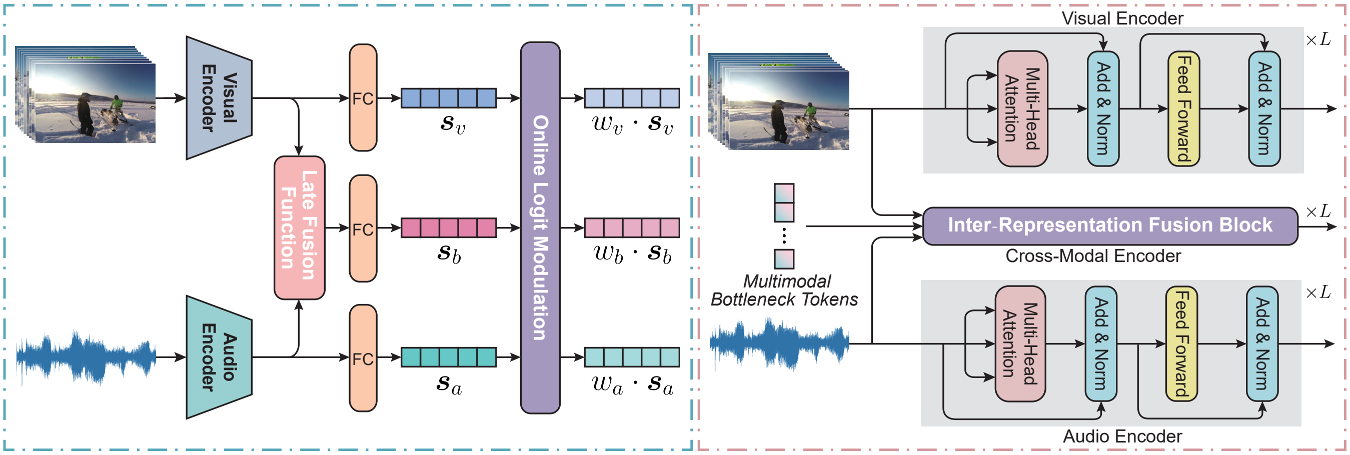


Figure 2: Take visual and audio modalities as an example. OLM learns model-agnostic dynamic *logit coefficients*, accommodating both late fusion and mid fusion paradigms. The left panel illustrates the generic paradigm of *late fusion*. Following the regulation of logit coefficients, the directions of the logit vectors for each modality remain unchanged, while the magnitudes of these vectors are harmonized to achieve better balance. The right panel demonstrates an exemplar of *mid fusion*. To decouple unimodal feature learning and cross-modal interaction learning for multimodal sequence inputs, we introduce an intermediate representation fusion block (see § 3.1 for architectural details.)

3 Method

Problem Definition and Notation

Empirically, we consider two modalities *e.g.*, *visual* and *audio* modality, as an exemplar to illustrate our core idea. Given a labeled dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{\mathsf{N}}$, where $y_i \in \{1, 2, \ldots, C\}$ represents the category associated with x_i. C is the number of classes and N is the number of samples. For each sample x_i, we extract visual feature sequence $\mathbf{z}_v^{[0]} \in \mathbb{R}^{\mathsf{T}_v \times d_v}$ and acoustic feature sequence $\mathbf{z}_a^{[0]} \in \mathbb{R}^{\mathsf{T}_a \times d_a}$, where $\{\mathsf{T}_m\}_{m\in\{v,a\}}$ is the sequence length and $\{d_m\}_{m\in\{v,a\}}$ is the feature dimension of each modality. Let $\mathbf{z}_m$ denote the sequence of token representations generated by the multimodal interactions. $Tf(\cdot)$ indicates the processing of Transformer layers (blocks). Our goal is to learn a robust model that can efficiently integrate all channels of multimodal information to predict y.

3.1 Multimodal Fusion Framework

Fig. 2 illustrates two distinct multimodal fusion frameworks, namely, *late fusion* and *mid fusion*. Subsequently, we provide an in-depth exposition of the proposed Intermediate Representation Fusion Block (IRFB).

Unimodal Encoder

Analogous to [Devlin *et al.*, 2018; Dosovitskiy *et al.*, 2020], we first append an additional learnable [CLS] token to each modality input, and then employ the standard Transformer [Vaswani *et al.*, 2017] as a unimodal encoder to obtain token embeddings for each modality at each layer:

$$\begin{aligned}\mathbf{z}_v^{[\ell+1]} &= Tf_1(\mathbf{z}_v^{[\ell]}),\\ \mathbf{z}_a^{[\ell+1]} &= Tf_2(\mathbf{z}_a^{[\ell]}),\end{aligned} \tag{1}$$

where ℓ indexes the layer number of the transformer models. Since the appended [CLS] token aggregates the information from all tokens, we use its embedding as the utterance-level representation for each modality.

Intermediate Representation Fusion Block

To effectively model the interactions among the utterance-level intermediate representations, we introduce IRFB (see Fig. 3). This block not only facilitates capturing interactions among the multimodal intermediate representations within the same layer but also enables capturing interactions among different layers. It is worth noting that directly fusing multiple modalities in a *one-to-one* manner can be inefficient, particularly when dealing with multiple modalities simultaneously [Sun *et al.*, 2023]. To overcome this inefficiency, we use a set of bottleneck tokens, denoted as $\mathbf{z}_b$, as a central message hub to facilitate communication with each modality, drawing inspiration from [Nagrani *et al.*, 2021]. The *multimodal bottleneck token* set $\mathbf{z}_b$ is randomly initialized and the number of tokens is set to T_b, *i.e.*, $\mathbf{z}_b = \{z_i\}_{i=1}^{\mathsf{T}_b}$. T_b is typically much smaller than T_v or T_a. To interact with unimodal features, we employ the multi-head attention mechanism [Vaswani *et al.*, 2017] as follows:

$$\begin{cases}\mathbf{z}_{v\to b} = \mathrm{LayerNorm}(\mathbf{z}_b + \mathrm{Att}(\mathbf{Q}_b^{(v)}, \mathbf{K}_v, \mathbf{V}_v)),\\ \mathbf{z}_{a\to b} = \mathrm{LayerNorm}(\mathbf{z}_b + \mathrm{Att}(\mathbf{Q}_b^{(a)}, \mathbf{K}_a, \mathbf{V}_a)),\end{cases} \tag{2}$$

where $\mathbf{Q}_b^m = \mathbf{z}_b \mathbf{W}_{bm}^Q, \mathbf{K}_m = \mathbf{z}_m \mathbf{W}_m^K$ and $\mathbf{V}_m = \mathbf{z}_m \mathbf{W}_m^V$ are linear transformations of the bottleneck tokens and unimodal input sequences, $m \in \{v, a\}$. To model interactions across different layers, we adopt a tanh-gating mechanism [Hochreiter and Schmidhuber, 1997], which effectively filters out irrelevant information while retaining valuable information flow by:

$$\begin{cases}\mathbf{g}_{v\to b}^{[\ell]} = \mathrm{Sigmoid}(\mathbf{W}_{vb}^{\ell}[\mathbf{z}_{v,\mathtt{cls}}^{[\ell]}, \mathbf{z}_{a,\mathtt{cls}}^{[\ell]}] + \boldsymbol{b}_{vb}^{\ell}),\\ \mathbf{g}_{a\to b}^{[\ell]} = \mathrm{Sigmoid}(\mathbf{W}_{ab}^{\ell}[\mathbf{z}_{v,\mathtt{cls}}^{[\ell]}, \mathbf{z}_{a,\mathtt{cls}}^{[\ell]}] + \boldsymbol{b}_{ab}^{\ell}),\\ \mathbf{z}_{v\to b}^{[\ell]} = \mathrm{LayerNorm}(\mathbf{z}_{v\to b}^{[\ell]} + \mathbf{g}_{v\to b}^{[\ell]} \odot \mathbf{z}_{v\to b}^{[\ell]}),\\ \mathbf{z}_{a\to b}^{[\ell]} = \mathrm{LayerNorm}(\mathbf{z}_{a\to b}^{[\ell]} + \mathbf{g}_{a\to b}^{[\ell]} \odot \mathbf{z}_{a\to b}^{[\ell]}),\end{cases} \tag{3}$$

where $[,]$ denotes the concatenation along the feature dimension, $\mathbf{W}_{mb}^{\ell} \in \mathbb{R}^{(d_v+d_a)\times d}$ are the layer-specific weight ma-

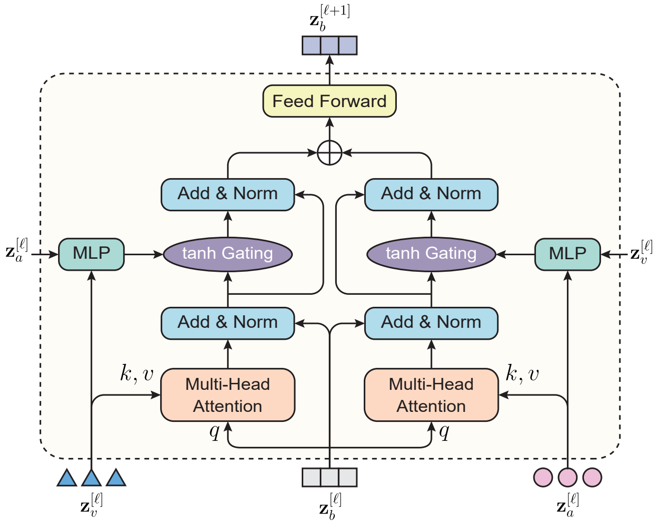


Figure 3: Illustration of the architectural details of IRFB.

trices, and $\boldsymbol{b}_{mb}^{\ell} \in \mathbb{R}^d$ is the bias and $m \in \{v, a\}$. Denote by $\mathbf{z}_g = \mathrm{Concat}(\mathbf{z}_{v\to b}^{[\ell]}, \mathbf{z}_{a\to b}^{[\ell]})$ the temporal concatenation of cross-attended feature sequences. To aggregate refined information from different modalities, the update rule of $\mathbf{z}_b$ is finally defined as:

$$\mathbf{z}_b^{[\ell+1]} = \mathbf{z}_b^{[\ell]} + \mathrm{Softmax}(\boldsymbol{v}^\top \mathtt{tanh}(\mathbf{z}_g \mathbf{W} + \boldsymbol{b})) \cdot \mathbf{z}_g, \qquad (4)$$

where $\boldsymbol{v}$, $\mathbf{W}$, and $\boldsymbol{b}$ are layer-specific learnable parameters. We omit the layer indexes of these parameters for brevity.

Remarks. A multimodal transformer equipped with IRFB effectively disentangles the process of unimodal feature acquisition from that of multimodal interactions, thus allowing for the decoupled learning of inter-modal features through modal fusion while ensuring sufficient and independent learning of unimodal features. Theoretically, our unimodal encoders can achieve the same performance as the unimodal baselines when trained long enough.

3.2 Online Logit Modulation

After the encoding of a multimodal transformer, we obtain a set of token representations $\{\mathbf{z}_v, \mathbf{z}_a, \mathbf{z}_b\}$ for the visual modality, audio modality, and fused modality, respectively. These token representations will be pooled along the temporal dimension and fed into three independent classifiers to generate the *logits* (also named *logit vectors*). That is, the *pre-softmax logits* can be calculated as the following:

$$\boldsymbol{s}_m = \boldsymbol{W}_m^\top \cdot \mathtt{SeqPooling}(\mathbf{z}_m), \forall m \in \{v, a, b\}, \qquad (5)$$

where $\boldsymbol{W}_m^\top \in \mathbb{R}^{d\times C}, m \in \{v, a, b\}$ denote the three separate linear classifiers. Here, the SeqPooling indicates the simple sequence pooling, where the average of the output vectors is taken as the summary representation.

Without loss of generality, a *logit vector* $\boldsymbol{s}_m$ can be decomposed as $\boldsymbol{s}_m = \|\boldsymbol{s}_m\|_2\, \hat{\boldsymbol{s}}_m$, where $\|\cdot\|$ denotes the L_2 norm and $\hat{\boldsymbol{s}}_m$ is the unit vector in the same direction as $\boldsymbol{s}_m$. In other words, $\|\boldsymbol{s}_m\|_2$ and $\hat{\boldsymbol{s}}_m$ indicate the *magnitude* and the *direction* of the logit vector $\boldsymbol{s}_m$, respectively. Previous studies have elucidated that distinct modalities exhibit varying levels of convergence and generalization rates [Wang *et al.*, 2020; Nagrani *et al.*, 2021]. Our objective is to devise a set of dynamic blending weights $\{w_m\}$, $\forall m \in \{v, a, b\}$ to improve joint multimodal training, which ensures similar rates of parameter updates across different modalities during training. To achieve this, we impose constraints on the magnitude of logit vectors corresponding to each modality. In order to determine the optimal magnitude value for each modality, we take into account both the model's convergence rate and generalization rate, similar to [Wang *et al.*, 2020]. Concretely, we gauge the generalization rate at the n-th iteration step using:

$$G_n = |\mathcal{L}_{dev}(\boldsymbol{\Theta}^{[0]}) - \mathcal{L}_{dev}(\boldsymbol{\Theta}^{[n]})|, \qquad (6)$$

where $\mathcal{L}_{dev}$ represents the validation loss, and $\boldsymbol{\Theta}$ denotes network parameters. Similarly, we measure the convergence rate C at the n-th iteration step as follows:

$$\begin{aligned} C_n = \big|&|\mathcal{L}_{dev}(\boldsymbol{\Theta}^{[n]}) - \mathcal{L}_{train}(\boldsymbol{\Theta}^{[n]})| \\ &- |\mathcal{L}_{dev}(\boldsymbol{\Theta}^{[0]}) - \mathcal{L}_{train}(\boldsymbol{\Theta}^{[0]})|\big|, \end{aligned} \qquad (7)$$

where $\mathcal{L}_{train}$ denotes the training loss. We then compute the generalization and convergence rates for each modality m, denoted as $\{G_{n,m}\}$ and $\{C_{n,m}\}$. A higher convergence rate indicates a higher risk of overfitting for a modality.We encourage the use of smaller logit magnitudes to counteract overfitting and larger logit magnitudes to mitigate underfitting. To this end, we can derive a set of modality-aware rebalancing factors by:

$$\{\lambda_{n,v}, \lambda_{n,a}, \lambda_{n,b}\} = \mathrm{Softmax}([\frac{G_{n,v}}{C_{n,v}}, \frac{G_{n,a}}{C_{n,a}}, \frac{G_{n,b}}{C_{n,b}}]). \qquad (8)$$

Building upon these modal rebalancing factors, we first outline a baseline for different modality logit magnitudes, which can be an average logit norm of weighted modality logit vectors: $\bar{s}_n = \mathbb{E}_{m\in\{v,a,b\}}(||w_{m,n} \cdot \boldsymbol{s}_{m,n}||_2)$. Then we construct the *target logit magnitude* for each modality via $\lambda_{n,m} \cdot s_n$. Our goal is to find an optimal set of blending logit coefficients that modulate the magnitude of the logit vector per modality to align the target logit magnitude. Therefore, the logit modulation loss can be formalized as follows:

$$\mathcal{L}_{logit} = \sum\nolimits_{m\in\{v,a,b\}} \big|w_m \cdot ||\boldsymbol{s}_m||_2 - \lambda_m \cdot \bar{s}\big|. \qquad (9)$$

Note that the optimization of Eq. 9 is nonlinear due to the L_2 norms involved in the expression. As a result, it is unlikely to find a global analytical solution for the optimal values of $\{w_v, w_a, w_b\}$. Instead, we employ the gradient descent algorithm to search for the optimal coefficients that lead to the desired minima. Algorithm 1 describes a pipeline for solving a set of *logit coefficients* using the SGD optimizer.

4 Experiment

4.1 Datasets and Metrics

Kinetics-Sounds (KS) [Arandjelovic and Zisserman, 2017] is a subset of 36 human action classes selected from the Kinetics dataset [Kay *et al.*, 2017], comprising 10-second videos sampled at 25fps from YouTube. In line with [Peng *et*

Algorithm 1: Online Logit Modulation (OLM)

```
Input  : Logit vectors $\{s_v, s_a, s_b\}$ output by unimodal and
         cross-modal encoders; the training loss
         $\mathcal{L}_{train}(m, n)$ and validation loss $\mathcal{L}_{dev}(m, n)$ for
         each modality $m$ at the $n^{th}$ iteration step; and the
         learning rate $\eta$
Output: Logit Coefficients $\{w_{v,n}, w_{a,n}, w_{b,n}\}$
Initialize the learning rate and the logit coefficients
  $\{w_v, w_a, w_b\}, w_v(0) \leftarrow 1, w_a(0) \leftarrow 1, w_b(0) \leftarrow 1$;
Compute the rebalancing fators $\{\lambda_v, \lambda_a, \lambda_b\}$ according to
  Eq. 8;
for $i \leftarrow 1$ to max_iteration_step do
    // compute an average $\ell_2$-norm of
       different logit vectors
    $\bar{s} \leftarrow \mathbb{E}_{m\in\{v,a,b\}}(||w_m \cdot s_m||_2)$;
    // compute the logit modulation loss
       between the current weighted logit
       norms and target ones
    $\mathcal{L}_{logit} \leftarrow$ Eq. 9;
    // update the coefficients using
       stochastic gradient descent
    $w_v(i) \leftarrow w_v(i-1) - \eta * \frac{\partial \mathcal{L}_{logit}}{\partial w_v}$;
    $w_a(i) \leftarrow w_a(i-1) - \eta * \frac{\partial \mathcal{L}_{logit}}{\partial w_a}$;
    $w_b(i) \leftarrow w_b(i-1) - \eta * \frac{\partial \mathcal{L}_{logit}}{\partial w_b}$;
Return: $\{w_{v,n}, w_{a,n}, w_{b,n}\} = \{w_v(i), w_a(i), w_b(i)\}$;
```

	VGGSound		CREMA-D		KS	
Method	**Acc.**	**mAP**	**Acc.**	**mAP**	**Acc.**	**mAP**
Unimodal Baselines						
Audio-only	44.3	48.4	52.5	54.2	55.2	57.4
Visual-only	31.0	34.3	41.9	43.0	43.5	45.8
Bimodal Fusion Baselines						
Concat$^\diamondsuit$	49.1	52.5	51.7	53.5	59.8	61.9
Sum	49.2	52.4	51.5	53.5	58.5	60.6
FiLM†	48.6	51.6	50.6	52.1	57.3	60.0
Gated‡	49.3	52.2	51.7	53.3	59.1	62.1
Attention	49.6	51.7	52.4	54.9	60.3	63.2
GradNorm$^\diamondsuit$	49.8	52.4	54.6	57.2	60.2	62.9
MMCosine$^\diamondsuit$	50.1	52.9	57.7	60.3	61.5	64.4
AVSlowFast	50.8	53.7	61.6	64.2	62.6	64.7
G-Blend$^\diamondsuit$	49.9	52.8	56.8	59.6	62.2	**65.7**
CUR	49.6	52.3	56.5	59.1	60.7	63.6
PMR$^\diamondsuit$	50.2	52.5	61.8	64.5	62.8	65.6
OGM-GE$^\diamondsuit$	50.6	53.9	61.9	63.9	62.3	65.2
OLM-Conv$^\diamondsuit$	**51.1**	**54.1**	**62.4**	**65.2**	**63.1**	65.5

Table 1: Performance Comparison across the VGGSound, CREMA-D, and KS datasets. Results marked by $\diamondsuit$ are obtained under the late fusion by concatenation. To ensure fair comparisons, apart from CUR and AVSlowFast, all approaches employ an identical encoder.

al., 2022], we focus on 31 action categories that can be recognized visually and auditorily, including actions like playing various musical instruments. This dataset contains 19,000 10-second video clips, with 15,000 clips used for training, 1,900 for validation, and 1,900 for testing.

VGGSound [Chen *et al.*, 2020a] is a large-scale video dataset comprising 309 classes, with nearly 200K 10-second video clips capturing a diverse range of audio events in everyday life. Each clip's sound source is visually presented in the video, demonstrating clear audio-visual correspondence. After filtering out unavailable videos, we obtained 168,618 videos for training and validation, and 13,954 for testing.

AVE [Tian *et al.*, 2018] is a subset of the AudioSet dataset [Gemmeke *et al.*, 2017] designed for audio-visual event localization. It comprises 28 event categories, consisting of 4,143 10-second videos. This dataset encompasses a diverse range of audio-visual events from various domains, with each video containing at least one 2-second long audio-visual event, annotated with frame-level boundaries. The training, validation, and test sets are divided into 3,339, 402, and 402 samples, respectively.

CREMA-D [Cao *et al.*, 2014] is a multimodal dataset designed for speech emotion recognition. It comprises 7,442 video clips, each lasting 2 to 3 seconds, featuring 91 actors delivering concise utterances. The dataset encompasses six of the most prevalent emotions: *anger*, *happiness*, *sadness*, *neutral*, *disgust*, and *fear*. It consists of 6,698 samples for training and validation, with 744 samples for testing.

MER-MULTI is a subchallenge of the MER2023 [Lian *et al.*, 2023], aiming to simultaneously recognize discrete emotions and valence in given raw video clips. The discrete emotion categories include *happiness*, *neutral*, *anger*, *sadness*, *worry*, and *surprise*. Valence is an emotional dimension with values ranging from -5 to 5, reflecting the degree of emotional pleasantness. This dataset comprises 3,373 video clips for training and 411 video clips for testing.

SUNRGBD V1 [Song *et al.*, 2015] comprises 10,335 RGB-D images collected from different sensors. We evaluate OLM on the scene classification task, which entails categorizing a given RGB-D image into one of the predefined 19 scene categories. We partition the data into training and testing sets, ensuring that approximately half of the data from each sensor goes into two subsets. Given that some images were captured from the same building or house with similar furniture styles, we ensure that images from the same building are either entirely within the training set or exclusively in the testing set.

The evaluation metrics employed in our experiments include the commonly used top-1 accuracy, F1-score, mean average precision (mAP), and mean squared error (MSE).

4.2 Experimental Settings

We evaluate two model variants. One variant is implemented with traditional CNN-based encoders combined with late fusion (refer to Fig. 2), termed OLM-Conv. The other variant uses transformer-based encoders equipped with the Intermediate Representation Fusion Module (IRFM), referred to as OLM-Trans. Both variants of our model employ the online logit modulation strategy. Specifically, for OLM-Conv, we use ResNet18 [He *et al.*, 2016] as the encoders following previous works [Zhao *et al.*, 2018; Peng *et al.*, 2022]. **AVE**, **Kinetics-Sounds**, and **VGGSound** datasets consist of videos with a duration of 10 seconds each. To process these videos, we extract frames at a rate of 1fps and uniformly sample 3 frames from each clip, which serve as the visual input for our model. For the audio data, we utilize a window of length 512

Network	**AVE Localization (Acc. %)**			
	Baseline	**G-Blend**	**OGM-GE**	`OLM-Conv`
AVGA [2018]	72.0	72.2	**72.8**	72.6
MAFnet [2021]	73.2	73.8	74.1	**74.8**
PSP [2021]	76.2	76.5	76.9	**77.3**
CMBS [2022]	79.3	80.1	81.2	**81.8**

Table 2: Performance Evaluation of the Audio-Visual Event Localization (AVE) task across different approaches in conjunction with diverse backbone network architectures.

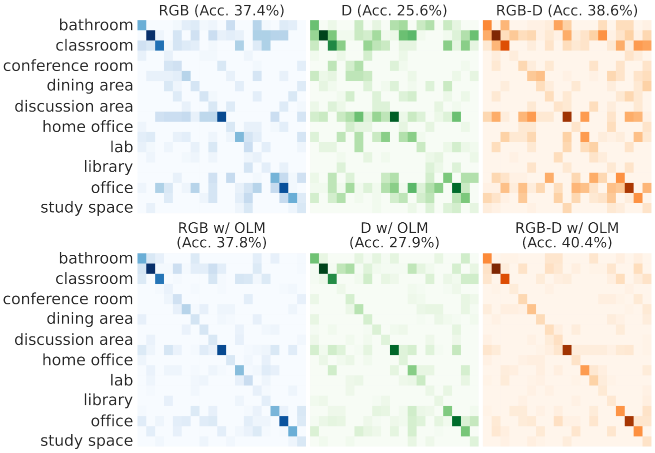


Figure 4: Confusion Matrix Comparison between models trained with and without `OLM` on the SUNRGBD dataset, employing Places-CNN as the feature extractor and late fusion by simple *concatenation*. The average classification accuracy across 19 categories under each modality (*e.g.*, RGB, D (depth), and RGB-D) is recorded.

with an overlap of 353 to transform the raw audio data into spectrograms of size 257 × 1004 using the librosa [McFee *et al.*, 2015] library. As for **SUNRGBD**, we adopt Places-CNN [Zhou *et al.*, 2014], which reaches the optimal performance for color-based scene classification on the SUN database [Xiao *et al.*, 2010], for feature extraction of both RGB and depth images. Regarding **CREMA-D**, its video clips last from 2 to 3 seconds. From each clip in CREMA-D, we extract 1 frame and use a window of length 512 with an overlap of 353 to convert the audio data into spectrograms of size 257 × 299. For **MER-MULTI**, we first extract human face images using the OpenFace toolkit. The pre-trained MANet [Zhao *et al.*, 2021], HuBERT [Hsu *et al.*, 2021], and MacBERT [Cui *et al.*, 2020] models were employed for the extraction of visual, audio, and textual features, respectively. For `OLM-Trans`, we stack six standard transformer blocks and `IRFB` blocks (cf. Sup. D for implementation details).

4.3 Comparison with State-of-the-Arts

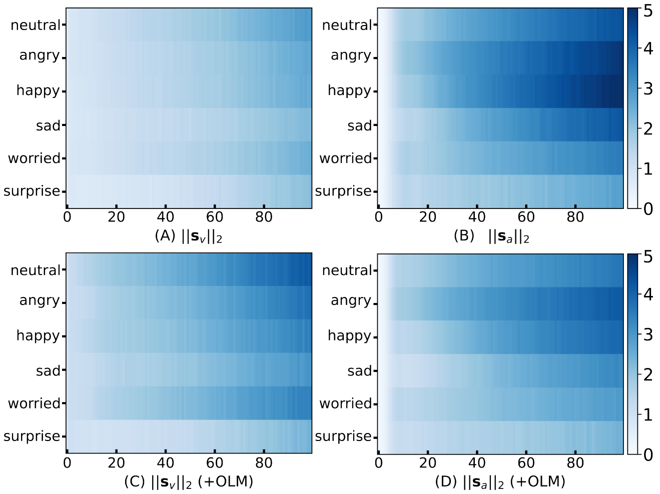
We compared `OLM` with a wide range of baseline methods, including two unimodal baselines and four simple bimodal baselines: *concatenation*, *summation*, *gated*, and *attention*. Additionally, we explored several advanced techniques for multimodal fusion, alignment, and training, encompassing:

- **FiLM** [Perez *et al.*, 2018] performs a simple feature-wise affine transformation on the intermediate features of a neural network based on conditional information.
- **GradNorm** [Chen *et al.*, 2018] proposes gradient normalization to balance training in deep multitask models by dynamically tuning gradient magnitudes.
- **MMCosine** [Xu *et al.*, 2023b] imposes modality-wise L_2 normalization to features and weights by cosine similarity towards balanced multi-modal learning.
- **AVSlowFast** [Xiao *et al.*, 2020] is a multimodal extension of SlowFast [Feichtenhofer *et al.*, 2019], which incorporates a faster audio pathway and deeply fuses audio and visual features at multiple levels.

Modality	**Attention**		`OLM-T-MBT`		`OLM-T-IRFB`	
	F1 (↑)	**E** (↓)	**F1** (↑)	**E** (↓)	**F1** (↑)	**E** (↓)
Unimodal Baselines						
Audio-only	65.7	1.27	65.7	1.27	65.7	1.27
Visual-only	57.5	1.38	57.5	1.38	57.5	1.38
Text-only	42.7	2.39	42.7	2.39	42.7	2.39
HOG-only	55.6	1.46	55.6	1.46	55.6	1.46
Bimodal Fusion Results						
A+T	67.1	1.16	71.2	0.89	73.8	0.84
A+V	73.2	0.86	77.9	0.68	**80.8**	**0.76**
V+T	61.2	1.28	65.3	1.21	68.6	1.21
A+H	72.7	0.88	77.3	0.69	80.6	0.77
Mulitmodal Fusion Results						
A+V+T	75.8	0.92	78.4	0.85	80.2	0.74
A+H+T	74.9	0.91	77.1	0.89	80.1	0.76
A+V+H	76.2	0.89	78.9	0.81	81.1	**0.70**
A+V+T+H	76.6	0.87	79.5	0.78	**81.4**	0.72

Table 3: Impact of `OLM` applied to different modalities and comparison regarding distinct *middle fusion* manners. Here, 'F1' denotes the F1-score, while 'E' signifies the mean squared error. '↑' indicates the higher values the better performance, while '↓' indicates the lower values the better performance. All methods share identical unimodal Transformer-based encoders and an attention fusion.

The remaining comparison methods for modulating the pace of multimodal training, including G-Blend [Wang *et al.*, 2020], CUR [Wu *et al.*, 2022], OGM-GE [Peng *et al.*, 2022] and PMR [Fan *et al.*, 2023], are introduced in § 2. Table 1 reveals several intriguing observations: *i*) a performance imbalance among modalities, with the audio modality exhibiting dominance. For instance, the performance of audio-only baselines significantly surpasses that of visual-only baselines across the three datasets; *ii*) Occasionally, the performance of unimodal baseline surpasses that of simple multimodal fusion baselines, as witnessed in the case of CREMA-D where the audio-only baseline outperforms all bimodal fusion baselines. This indicates potential under-optimization of multimodal models due to naive multimodal joint training; *iii*) The proposed `OLM` strategy exhibits advantages compared to the other competitors. Specifically, `OLM-Conv` attains superior or comparable accuracy and mAP scores across the three datasets. To further validate the versatility of our strategy, we apply `OLM` to the AVE localization task. Table 2 compares `OLM-Conv` with several competitive counterparts on this task. As seen, when combined with diverse AVE backbone networks, `OLM-Conv` yields the most significant improvement over the baseline in most cases, confirming the



(A) $||\mathbf{s}_v||_2$

(B) $||\mathbf{s}_a||_2$

(C) $||\mathbf{s}_v||_2$ (+OLM)

(D) $||\mathbf{s}_a||_2$ (+OLM)

Figure 5: Visualization of L_2 norms of logit vectors corresponding to each class for both visual and audio modalities over increasing training epochs. Figures (A) and (B) show the logit norms without OLM, while figures (C) and (D) depict logit norms with OLM applied.

effectiveness of logit modulation in bolstering multimodal training. Worth noting is that unlike G-Blend and OGM-GE, our OLM neither necessitates additional unimodal networks to compute per-modality weights nor involves intricate gradient calibrations, therefore boasting greater out-of-the-box applicability. Furthermore, we present the confusion matrices generated by models with and without the OLM strategy applied in Fig. 4. The OLM-trained model exhibits pronounced enhancement over the non-OLM-trained model (*w.r.t. acc. and recall*), highlighting the OLM's generalization ability.

4.4 Ablation Study

Impact of Different Modalities and IRFB

Table 3 presents the unimodal baseline results and attention fusion-based multimodal baseline results on MER-MULTI. Particularly, we investigate two variants of our model, namely OLM-T-MBT and OLM-T-IRFB. The former employs a non-decoupled multimodal attention bottleneck (MBT) [Nagrani *et al.*, 2021] for cross-modal mid fusion, while the latter employs IRFB for mid fusion. Both model variants share identical transformer-based unimodal encoders and use the same attention late fusion to generate unimodal and cross-modal outputs. In Table 3, several observations merit attention: a) Disparities in performance among different modalities are evident. The text modality performs worst, while the audio modality performs best; b) OLM substantially boosts the performance of any modality combination baseline, showcasing its versatility; c) OMT-T-IRFB outperforms OMT-T-MBT, with an average improvement of approximately 2%.

Visualization of Logit Modulation

Fig. 5 illustrates the class-specific L_2 norm of logit vectors for the visual and audio modalities. It is evident that, as training progresses, the magnitude of logit vectors per class continues to increase. However, the rate of logit norm growth varies across modalities, with the logit norm for the audio modality increasing more rapidly than that for the visual modality.

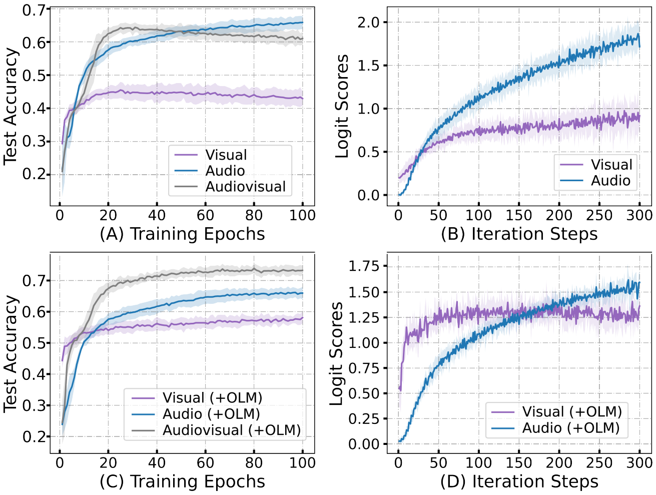


Figure 6: (A) Test accuracy of unimodal and audiovisual fusion via concatenation. (B) Evolution of the batch-averaged logit scores corresponding to each class for visual and audio modalities during training. (C) Test accuracy of unimodal and audiovisual with OLM applied. (D) The batch-averaged unimodal logit scores with OLM.

This suggests a faster learning pace for audio modality features and potential underfitting in the visual modality. OLM operates on the logit magnitudes of diverse modalities. Analysis of Fig. 5 (C) and (D) reveals that the rate of logit norm growth in the audio modality decelerates, whereas it accelerates in the visual modality, thus achieving modality-aware balanced training. The same trend is also discernible from the logit score curves in Fig. 6 (B) and (D). Fig. 6 (A) and (B) aptly demonstrate that OLM mitigates the underfitting issue of visual modality in the context of multimodal joint training. These figures intuitively underscore the benefits of OLM in the realm of multimodal training: the ability to foster more robust multimodal features while ensuring comprehensive unimodal feature training, thereby unlocking the genuine potential of multimodal models.

5 Conclusion

In this work, we present two core components, IRFB and OLM, tailored to enhance multimodal training. Specifically, IRFB efficiently alleviates potential underfitting concerns of unimodal networks in imbalanced multimodal learning by disentangling the learning of unimodal features from multimodal interactions. OLM, unrestricted by model architectures and fusion methods, mitigates the suppression of dominant modalities on other modalities. This is achieved by regulating the magnitude of the logit vector for each modality, aligning it with its modality-aware target. Empirical studies have fully unveiled the efficacy of IRFB and the superiority of OLM over prevailing multimodal training calibration alternatives.

Contribution Statement

Daoming Zong and Chaoyue Ding contributed equally to this paper.

References

[Alvarez-Melis and Fusi, 2020] David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *NeurIPS*, 33:21428–21439, 2020.

[Arandjelovic and Zisserman, 2017] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617, 2017.

[Brousmiche *et al.*, 2021] Mathilde Brousmiche, Jean Rouat, and Stéphane Dupont. Multi-level attention fusion network for audio-visual event recognition. *arXiv e-prints*, pages arXiv–2106, 2021.

[Cao *et al.*, 2014] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014.

[Chen *et al.*, 2018] Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *ICML*, pages 794–803. PMLR, 2018.

[Chen *et al.*, 2020a] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE, 2020.

[Chen *et al.*, 2020b] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607. PMLR, 2020.

[Cui *et al.*, 2020] Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Revisiting pre-trained models for chinese natural language processing. In *Findings of EMNLP*, pages 657–668, 2020.

[Devlin *et al.*, 2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Fan *et al.*, 2023] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *CVPR*, pages 20029–20038, 2023.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *ICCV*, pages 6202–6211, 2019.

[Gemmeke *et al.*, 2017] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *ICASSP*, pages 776–780. IEEE, 2017.

[Gretton *et al.*, 2012] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[Hazarika *et al.*, 2020] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *ACM MM*, pages 1122–1131, 2020.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

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

[Hsu *et al.*, 2021] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

[Huang *et al.*, 2022] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *ICML*, pages 9226–9259. PMLR, 2022.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021.

[Kay *et al.*, 2017] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *NeurIPS*, 34:9694–9705, 2021.

[Li *et al.*, 2022] Dongxu Li, Junnan Li, Hongdong Li, Juan Carlos Niebles, and Steven CH Hoi. Align and prompt: Video-and-language pre-training with entity prompts. In *CVPR*, pages 4953–4963, 2022.

[Lian *et al.*, 2023] Zheng Lian, Haiyang Sun, Licai Sun, Jinming Zhao, Ye Liu, Bin Liu, Jiangyan Yi, Meng Wang, Erik Cambria, Guoying Zhao, et al. Mer 2023: Multi-label learning, modality robustness, and semi-supervised learning. *arXiv preprint arXiv:2304.08981*, 2023.

[Lu *et al.*, 2019] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*, 32, 2019.

[McFee *et al.*, 2015] Brian McFee, Colin Raffel, Dawen Liang, Daniel P Ellis, Matt McVicar, Eric Battenberg, and

Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25, 2015.

[Nagrani *et al.*, 2021] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *NeurIPS*, 34:14200–14213, 2021.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, pages 8238–8247, 2022.

[Perez *et al.*, 2018] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, volume 32, 2018.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021.

[Shen *et al.*, 2023] Junhong Shen, Liam Li, Lucio M Dery, Corey Staten, Mikhail Khodak, Graham Neubig, and Ameet Talwalkar. Cross-modal fine-tuning: Align then refine. *arXiv preprint arXiv:2302.05738*, 2023.

[Song *et al.*, 2015] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015.

[Sun *et al.*, 2023] Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2023.

[Tian *et al.*, 2018] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, pages 247–263, 2018.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

[Wang *et al.*, 2020] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *CVPR*, pages 12695–12705, 2020.

[Wu *et al.*, 2022] Nan Wu, Stanislaw Jastrzebski, Kyunghyun Cho, and Krzysztof J Geras. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In *ICML*, pages 24043–24055. PMLR, 2022.

[Xia and Zhao, 2022] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *CVPR*, pages 19989–19998, 2022.

[Xiao *et al.*, 2010] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010.

[Xiao *et al.*, 2020] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[Xu *et al.*, 2023a] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE TPAMI*, 2023.

[Xu *et al.*, 2023b] Ruize Xu, Ruoxuan Feng, Shi-Xiong Zhang, and Di Hu. Mmcosine: Multi-modal cosine loss towards balanced audio-visual fine-grained learning. In *ICASSP*, pages 1–5. IEEE, 2023.

[Yang *et al.*, 2021] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *ICCV*, pages 11562–11572, 2021.

[Zhao *et al.*, 2018] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *ECCV*, pages 570–586, 2018.

[Zhao *et al.*, 2021] Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Transactions on Image Processing*, 30:6544–6556, 2021.

[Zhou *et al.*, 2014] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *NeurIPS*, 27, 2014.

[Zhou *et al.*, 2021] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *CVPR*, pages 8436–8444, 2021.