AMST: Alternating Multimodal Skip Training

Hugo Manuel Alves Henriques e Silva(✉)[1], Hongguang Chen[1], and Selpi[1]

Chalmers University of Technology and University of Gothenburg,
SE-41296 Gothenburg, Sweden {hugoalv,chenhon,selpi}@chalmers.se

Abstract. Multimodal Learning is one of the many fields in Machine Learning where models leverage the combination of various modalities to enhance learning outcomes. However, modalities may differ in data representation and complexity, which can lead to learning imbalances during the training process. The time it takes for a certain modality to converge during training is a crucial metric to determine modality imbalance. Given differences in convergence rates, different modalities may harmfully interfere with each other's learning process when simultaneously trained, as is commonly done in a multimodal scenario. To mitigate this negative impact, we propose Alternating Multimodal Skip Training (AMST) where the training frequency is adjusted for each specific modality. This novel method not only improves performance in conventional multimodal models that learn with fused modalities but also enhances alternating models that train each modality separately. Additionally, it outperforms state-of-the-art models while reducing training times.

Keywords: Multimodal Learning · Modality Imbalance · Modality Convergence Rate · Modality Bias · Skip Training · Training Frequency

1 Introduction

Multimodal Machine Learning aims to mimic how humans rely on different senses to analyse, judge, and act in distinct situations. Inspired by this aspect of human perception, multimodal models leverage diverse sensory inputs to improve learning outcomes. However, with multimodal learning comes its challenges. A key difficulty in recent multimodal approaches is the integration of information from multiple modalities, without having them negatively interfere with each other, in a way where rich modality-specific information is lost. Studies such as [14,15] have previously confirmed that some modalities tend to be more "dominant" than others, which causes the learning process to focus disproportionately on them, thereby limiting the contribution of other, potentially richer, modalities. Even though the "dominance" of one modality over another does not measure its informativeness or discriminative power, it affects the overall model performance.

Generally, raw data produced from different sensors and the way it is represented may differ in complexity when processing it. Consequently, more complex data may require longer training times than data with simpler representations. For example, when representing a video in both audio and visual modalities, it is

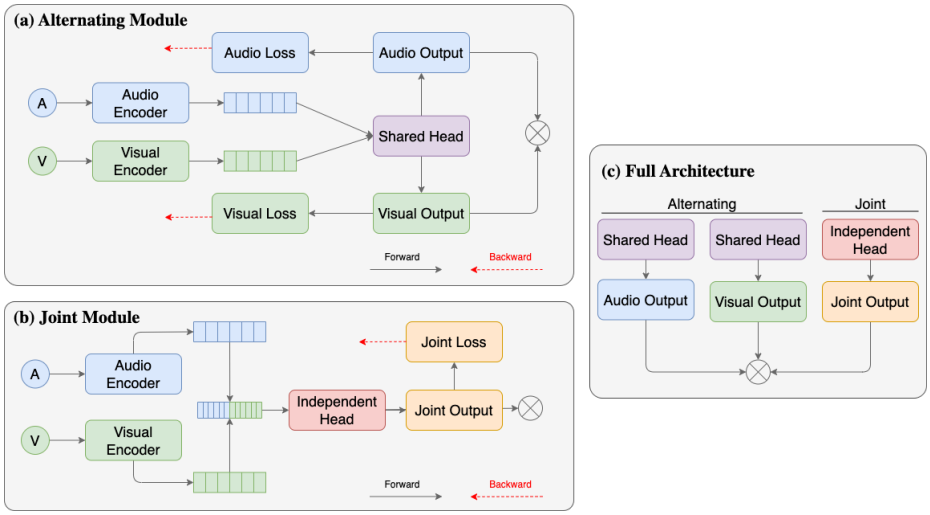

Fig. 1: Example of AMST's pipeline for 2 modalities (Audio and Visual): The Alternating Module **(a)** comprises both audio and visual encoders. The Shared Head predicts for audio and visual modalities individually. The Joint Module **(b)** consists of audio and visual encoders that produce a fused concatenated representation. The Independent Head predicts with the joint representation. The Full Architecture **(c)** averages the visual and audio outputs from the Alternating Module and the joint output from the Joint Module for a prediction.

relatively straightforward to create a single representation for the entire audio of the video. In contrast, generating a single representation for the visual modality often requires concatenating frames from various time steps, resulting in a more complex data representation overall. Recent methods do not address the issue of modality imbalance related to convergence rates, which leads to inefficient training and negatively affects the overall model performance. Neglecting differences in modality convergence rates allows certain modalities to overfit at different stages, resulting in the under optimisation of those that learn more slowly. When a modality overfits to the training data earlier than others, the multimodal model tends to prioritise it, causing premature convergence.

We address the problem of modality dominance related to convergence rates, by allowing the model to train more frequently on the data with the highest complexity, i.e., the slowest learning modalities, while periodically training it on the fastest learning ones. Inspired by human perception, the proposed method emulates how people allocate more effort to a more challenging task while still occasionally practising the simpler one. This uneven focus ultimately yields comparable performance across all tasks despite their different complexities.

Our approach, as illustrated in Figure 1, is built upon two modules, **Alternating** and **Joint**, both including skip training (see Figure 2). This approach led

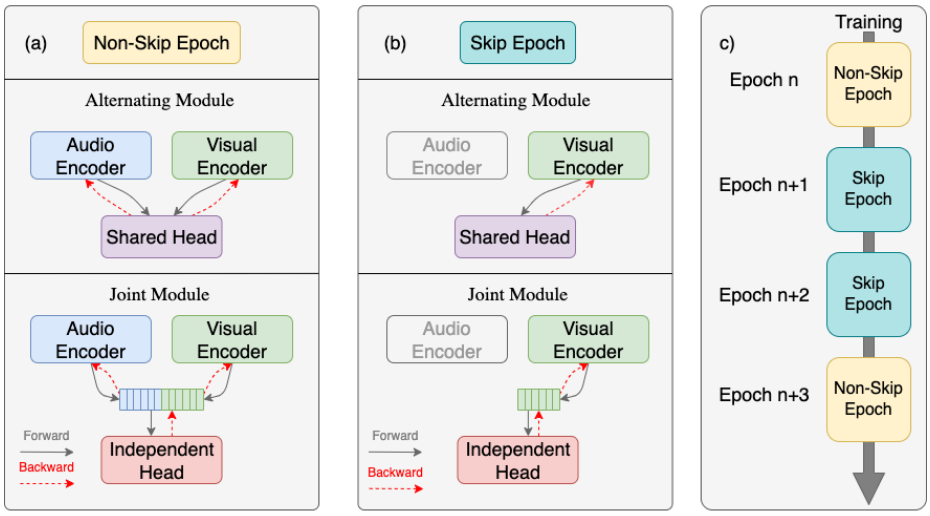

Fig. 2: AMST's training framework with skip parameter set to 3 for the audio modality and 1 for visual. **(a)** represents an epoch in which audio is not skipped and both modalities are updated. **(b)** represents an epoch in which the audio is skipped and only visual is updated. **(c)** represents how the training should proceed: 2 consecutive skip epochs as in **(b)** for every non-skip epoch as in **(a)**.

to the creation of the **A**lternating **M**ultimodal **S**kip **T**raining (**AMST**) method. The contribution of this paper is summarised as follows:

- We introduce Alternating Multimodal Skip Training (AMST), a novel method that (1) optimises the multimodal learning process without compromising the modality-specific information, and (2) addresses modality imbalance due to different convergence rates across modalities.
- We conduct experiments to show that AMST is able to (1) decrease the dominance gap related to convergence rates between the different modalities through skip training, (2) outperform the closest state-of-the-art methods, (3) increase the performance compared to when any of its key components are removed, and (4) greatly reduce overall computational expenses compared to the fastest existing methods.

The remainder of this paper is structured as follows: Section 2 introduces the related work. Our method AMST is presented in Section 3. The experiments are detailed in Section 4 and the results are discussed in Section 5. We conclude our work in Section 6. We provide the code at `https://github.com/CXianRen/AMST`.

2 Related Work

Modality convergence rate has previously been studied in works such as [14,15,7]. In [14], the authors demonstrated that different modalities overfit and generalise

at different rates, which makes conventional joint training of multimodal data not optimal, compared to the best unimodal counterpart. The finding that joint training of multimodal data is not optimal was also demonstrated in [12].

Several methods have been proposed to mitigate the impact of different convergence rates on the learning process. [11] proposed "on-the-fly gradient modulation", where the authors dynamically adjusted the gradient contributions of each modality during training. Inspired by [11], [6] developed an adaptive gradient modulation. Similarly, [7] took a comparable approach, while [3] introduced Prototypical Modal Rebalance (PMR). Here, prototypes were created. These represent centroids in the feature space for each class and serve as targets for each modality to cluster around, encouraging the alignment of slow-learning modalities. In [14], the authors proposed the gradient blending method to compute "an optimal blending of modalities based on their overfitting behaviours". Further, in [15], a metric called conditional learning speed was proposed to measure the learning speed of an individual modality relative to the other modalities. [12] suggested balancing the learning rates across modalities such that the modality nearing convergence would be given a lower learning rate so other modalities could catch up. The idea of adapting the learning rate was also proposed by [17].

In addition to balancing the learning rate, [12] also suggested separating the unimodal network from the multimodal network to allow modality-specific optimisation. Modality-specific optimisation is also promoted in a recent method called MLA [18] with its alternating approach. The aim of MLA is to provide each modality with an independent learning process in which they can fully leverage their information, without being disturbed by any other modality. Furthermore, cross-modal information is captured through a shared head between all modalities.

Our work builds on the alternating approach in MLA [18]. Despite MLA's efforts to minimise modality imbalance, we hypothesise that, similar to conventional multimodal joint training approaches, the model may remain biased toward the faster-learning modality, potentially compromising overall performance.

3 Alternating Multimodal Skip Training (AMST)

This section presents our proposed method. We describe both implemented modules, **Alternating** and **Joint**, explaining how they complement each other and how skip training is established in each of them, as displayed in Figure 2.

The Alternating Module, Figure 1 (a), aims to reframe the conventional multimodal joint training practice, where modalities are first fused and trained together. With the introduction of the alternating framework in [18], we allow each modality to train independently and learn a complete representation from its data by having modality-specific losses. Despite the training being done independently, we rely on a single head that is shared among all modalities. This shared head, which is simply a classification layer, aims to capture the cross-modal information while minimising the learning disturbance caused by the modalities' interaction.

The pipeline in the Alternating Module flows as shown in Figure 1 (a). A dedicated encoder is designed for each modality to generate its distinct latent representation. These representations are then processed by the shared head where different predictions are obtained per modality. For each modality, the model's parameters are optimised according to its modality-specific loss. Even though the training of a modality does not influence the encoders of other modalities, the shared head is always optimised, allowing minimal inter-modal communication.

Contrary to the Alternating Module, the Joint Module, Figure 1 (b), implements the conventional joint training practice, where full communication between the modalities is allowed, fusing them at a much earlier stage. However, scenarios that adopt this approach are more susceptible to modality imbalance related to convergence rates. This occurs because a single loss function is optimised, which inherently favours the faster learning modalities. As these modalities overfit and achieve near-perfect training performance, the overall model also attains near-perfect performance on the training data. Consequently, the premature convergence of the loss to near-zero values diminishes the effectiveness of gradient updates, hindering the learning of slower-learning modalities.

As displayed in Figure 1 (b), the Joint Module has its own encoders, which follow the same encoder structure as the Alternating Module. Each modality-specific encoder generates latent embeddings, which are subsequently fused into a single representation via concatenation. This fused representation is then processed by a classification layer, referred to as the independent head, as it operates separately from the Alternating Module's classification head.

3.1 The Motivation for Skip Training

We observed that each modality learns at different rates, converging at distinct time steps. Figure 3's (a) and (c) show the training accuracies of audio and visual modalities and the overall prediction on the CREMA-D dataset [2], for the Alternating and Joint Modules, respectively. In both plots, the audio and overall accuracy curves are closely aligned with each other, suggesting a bias towards the audio modality in both modules.

In the Alternating Module, Figure 3 (a), by epoch 25, the audio modality achieves near-perfect accuracy, whereas the visual modality remains at 40%. However, the overall prediction, derived from averaging the logits of both modalities' outputs from the shared head, closely follows the audio accuracy curve, suggesting a bias toward the audio modality in this module. Nevertheless, the visual modality continues to learn, albeit at a slower pace. Since rapid convergence of the audio modality does not hinder the learning of the visual modality in this module, we further examine the impact of not balancing modality convergence rates in an alternating setup. Figure 4 presents the average entropy values of each modality's predictions for each class in the CREMA-D dataset in different experiments. Lower entropy values indicate greater confidence in the corresponding modality's predictions. Figure 4 (a) shows the entropy values of both modalities in all classes in the default scenario (i.e., when no measures are applied to mitigate imbalances caused by differing convergence rates). One can see that the observed

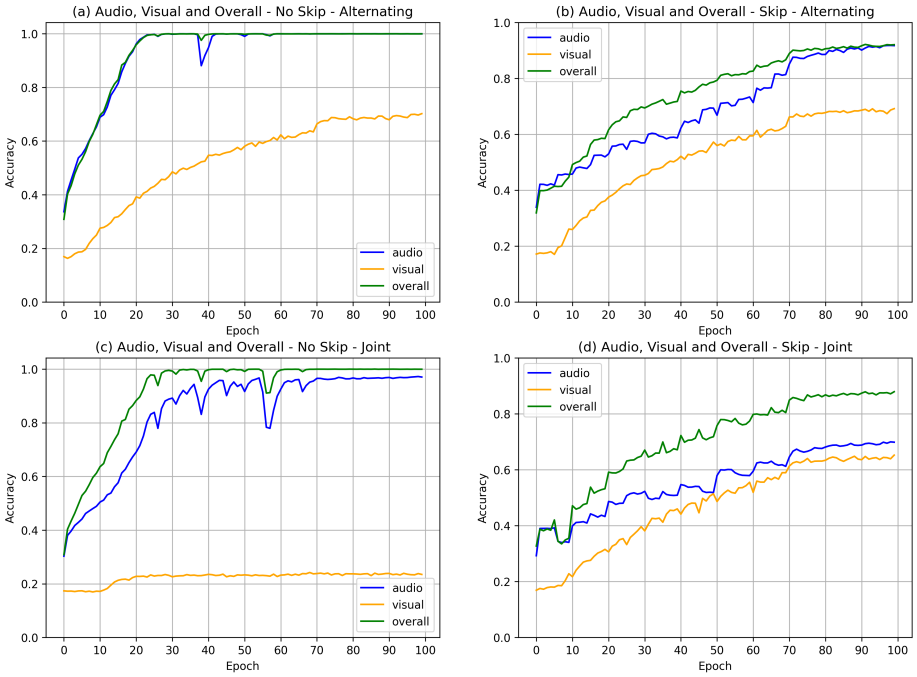

Fig. 3: Accuracies, for both Alternating and Joint modules, on CREMA-D dataset for audio and visual modalities as well as the overall case considering both modalities: **(a)** Alternating Module without skip training. **(b)** Alternating Module including skip training. **(c)** Joint Module without skip training. **(d)** Joint Module including skip training. The skip parameter for audio is 5 and for visual is 1.

entropy values for the audio modality are considerably lower than the visual ones, demonstrating the model's overconfidence in the audio modality. This further reinforces the notion that a bias is introduced in favour of the audio modality.

A potential way to mitigate overtraining of faster-learning modalities is the use of the early stopping technique (for neural networks) for a modality's training. However, as demonstrated in Figure 4 (b), where the training of the audio modality is halted around epoch 30 (5–10 epochs after approaching 100% accuracy), the gap in average entropy values observed in Figure 4 (a) remains. This indicates that entirely discontinuing a modality's training shortly after convergence does not eliminate the bias formed in its favour. Furthermore, these findings suggest that the early stages of training play a critical role in establishing this bias towards the faster-learning modalities. Thus, early stopping alone is insufficient to address the imbalance due to differing convergence rates.

For the Joint Module, Plot (c) in Figure 3 shows a pattern similar to the Alternating Module, where the overall joint accuracy closely follows the audio accuracy. By epoch 30, the overall accuracy reaches a near-perfect 100%, once

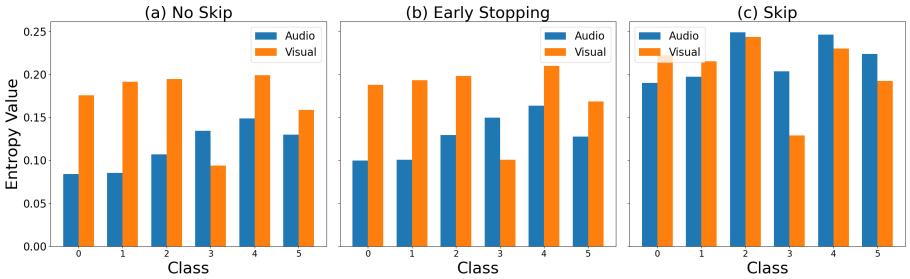

Fig. 4: Average entropy values for each class in the CREMA-D Dataset for 3 different methods in the Alternating Module. **(a)** - default method without skipping; **(b)** - early stopping. **(c)** - AMST's skip training. Each plot corresponds to the results of the best-performing model over 100 epochs of training.

again indicating a bias toward the audio modality due to its faster convergence. Yet, unlike in the Alternating Module, the visual accuracy stagnates as soon as the overall accuracy attains this near-perfect value. This occurs because, with a joint loss, as the loss approaches zero, gradient updates become progressively smaller, eventually reaching a point of insignificance. As a result, the visual modality fails to learn due to the absence of meaningful gradient updates.

Failing to address the bias in both scenarios can reduce overall model performance. To mitigate this, skip training is proposed as a solution to counteract the bias introduced by faster-learning modalities and their rapid convergence.

3.2 Skip Training in the Alternating Module

In this module, despite only the final layer being shared across modalities, given the learning curves observed in Figure 3 (a) and the substantial gap in average entropy values across modalities shown in Figure 4 (a), we hypothesised that a bias could develop in favour of the fastest-learning modalities. Were this to be the case, it would hinder the performance of the slow-learning ones, limiting their contribution and causing the model to overfit to the faster-learning counterparts.

Therefore, Skip Training is introduced in the Alternating Module. As shown in Figure 2 (b), the approach prioritises the learning process for slow-learning modalities, by skipping the optimisation of the faster-learning ones. This means that, during training, we determine with the help of an integer hyperparameter per modality, whether or not each modality should be optimised in the current epoch. Despite needing fine-tuning, a reasonable initial estimate for this integer hyperparameter can be obtained by analysing the rate of change in accuracy across all modalities during the initial epochs. Specifically, the skip parameter for each modality should be assigned in proportion to its accuracy rate of change relative to the slowest learning modality. For example, if a modality demonstrates an accuracy rate of change three times greater than that of the slowest learning modality in the early stages, its initial skip parameter should be set to three.

Algorithm 1 Alternating Skip Method

Input: Skip hyperparameters $s_1, \ldots, s_m$, Total epochs E, Modalities M
// **Training Stage**
for $e = 1$ to E **do** ▷ Iterate over epochs
  for each modality m in M **do** ▷ Iterate over modalities
    if $e \mod s_m = 0$ **then** ▷ Check skip schedule
      out := predModality(encode(m)) ▷ Generate prediction for modality m
      Backprop(Loss(out)) ▷ Backpropagate
    end if
  end for
end for
return trained_model
// **Inference Stage**
Init predictions $P \leftarrow []$ ▷ List to store modality predictions
for each modality m in M **do** ▷ Iterate over modalities for inference
  out := predModality(encode(m)) ▷ Generate prediction for modality m
  Append *out* to P ▷ Store prediction
end for
return Softmax(Average(P)) ▷ Compute final prediction by averaging

Given a modality M with respective integer skip parameter s_m, and current training epoch e, the decision to train M follows Equation (1). Algorithm 1 presents the pseudo-code for the alternating skip method mechanism. For inference, the default no-skip approach is used.

$$\text{Train}(M) = \begin{cases} True, & \text{if } e \bmod s_m = 0, \\ False, & \text{otherwise.} \end{cases} \tag{1}$$

This approach not only mitigates the overfitting of faster-learning modalities but also allows sufficient time for the model to learn from slower-learning ones. Furthermore, it alleviates some of the influence that the dominant modalities have on the updates of the shared head's parameters. During inference, the Alternating Module produces logit outputs for each modality which are then averaged to generate the final prediction.

To demonstrate the importance of having skip training in the alternating module, we compare Plots (b) and (a) of Figure 3, which illustrate the learning curves for the audio, visual, and overall module accuracies for the Alternating Module in the CREMA-D dataset, with and without skip training, respectively. As shown in Figure 3 (b), the audio and the overall accuracy curves exhibit a more iterative path compared to Figure 3 (a), while both plots' visual modality curves follow a similar trend. By epoch 70 in Figure 3 (b), all modalities and the overall accuracies begin to converge. This indicates a more balanced learning scenario, where modalities begin to converge at similar time steps, enabling the models to better utilise the multimodal setting. Furthermore, in Figure 4, we observe a significant reduction in the gap between the average entropy values of

the audio and visual modalities in (c) compared to (a) and (b). This indicates that the model has become nearly equally confident in both modalities. These results strongly support our initial hypothesis that there was a bias towards the faster-learning modality and demonstrate that skip training effectively mitigates this bias.

3.3 Skip Training in the Joint Module

To address the issue of modality imbalance related to convergence rate in joint architectures, skip training is applied similarly to how it is used in the Alternating Module. Figure 2, (b) illustrates the joint skip method. The same hyperparameters that define how many epochs each modality's optimisation is skipped during the alternating phase are reused to govern the joint skip method. Moreover, the initial skip parameter estimates from the Alternating Module apply similarly here. Like the Alternating Module, given a modality M with an associated integer skip parameter s_m and the current training epoch e, M is trained according to Equation (1).

A standard joint training approach concatenates modalities' embeddings in a fixed order. As defined in Equation (1), in skip training, a modality's latent representation is only used for prediction if its skip parameter divides the current training epoch. As a result, only the weights of the independent head and the encoders corresponding to these modalities are updated. At each epoch, a sub-head is derived from the independent head, retaining the weights of the modalities that are not skipped. The non-skipped modalities are then concatenated and fed into this sub-head to compute the final joint loss for that training epoch. This approach ensures that only the encoders of the non-skipped modalities, along with the corresponding weights of the independent head, are updated.

Figure 3 (d) illustrates the learning curves for the audio, visual, and overall module accuracies for the Joint Module in the CREMA-D dataset, incorporating skip training. With skip training, the audio modality and the overall accuracies undergo a more gradual learning process, allowing the visual modality to learn, since the overall joint loss does not reach zero in the early training stages. As in the Alternating Module, all modalities and the overall accuracies begin to converge around epoch 70.

During inference, the joint representation is formed by concatenating the latent embeddings of all modalities and feeding them into the independent head to generate the final prediction. Algorithm 2 presents the pseudo-code for the joint skip method.

3.4 Complementary Module Interaction

The Alternating and Joint Modules follow two different approaches. The former minimises inter-modal communication, while the latter maximises it. Both approaches have their advantages and drawbacks depending on the specific application. Notably, they operate independently, and skip training is not reliant on their co-existence, as each module is trained separately. By integrating skip

Algorithm 2 Joint Skip Method

Input: Skip hyperparameters $s_1, \ldots, s_m$, Total epochs E, Modalities M
// **Training Stage**
for $e = 1$ to E **do** ▷ Iterate over epochs
    Init $R \leftarrow []$ ▷ Store embeddings of active modalities
    Init active_modalities $\leftarrow []$ ▷ Track which modalities are active
    for each modality m in M **do** ▷ Determine active modalities
        if $e \mod s_m = 0$ **then** ▷ Check skip condition
            Append encodeModality(m) to R
            Append m to active_modalities
        end if
    end for
    Init reduced_head $\leftarrow$ Weights(independent_head, active_modalities)
    joint_representation := Concatenate(R) ▷ Form reduced joint representation
    out := reduced_head(joint_representation) ▷ Generate prediction
    Backprop(Loss(out)) ▷ Backpropagate
end for
return: trained_model
// **Inference Stage**
joint_representation := Concatenate({encodeModality(m) **for** $m \in M$})
return: independent_head(joint_representation) ▷ Generate final prediction

training with the full communication benefits of the joint approach and the independent training structure of the alternating method, the resulting fusion can generate more confident and aligned predictions. With skip training's reduced computational costs, employing both modules concurrently may become a feasible option depending on the task. When both modules are used simultaneously, the final prediction corresponds to a mean of the logit outputs of each modality from the Alternating Module and the logit output from the Joint Module.

4 Experiments

4.1 Data

To assess the performance of the proposed method, we utilised two audio-visual datasets: **CREMA-D** [2], **AVE** [13]; one visual-text dataset: **MVSA** [10]; and two audio-visual-text datasets: **IEMOCAP** [1] and **UR-FUNNY** [4].

CREMA-D [2] consists of 7,442 original clips displaying one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad). The AVE dataset [13] comprises 4143 10-second video clips of common actions and events. The dataset contains 28 different classes. MVSA-Single [10] is a multimodal sentiment analysis dataset that comprises 5129 image-text pairs collected from Twitter, labelled in 3 classes (positive, neutral and negative). The IEMOCAP dataset [1] comprises approximately 12 hours of audiovisual data, including video, speech, facial motion capture, and text transcriptions, labelled in angry, excited, frustrated, neutral

and sad. Finally, UR-FUNNY [4] consists of 16514 video segments with 8257 labelled as humorous and 8257 as non-humorous.

4.2 Experimental Setup

For AVE, the data was split into train, test, and validation sets following [13]. For the rest of the datasets, we randomly allocate 80% of the data for training, 10% for the validation, and 10% for the testing as no previous split is given. The MVSA-single dataset was processed as in [16]. ResNet18 [5] was used as the backbone encoder for both audio and visual modalities. For the text modality, RoBERTa [8] was used as the encoder with five unfrozen encoder layers. For the visual modality, for all datasets we extracted one frame per second from each video and selected three images as done in [18]. For MVSA, the single provided image per sample was used as the visual modality. For audio, we converted the raw audio samples into fbank (filterbank) [9] features. Regarding text, we tokenised the raw sample text with the RoBERTa tokeniser. To ensure complete fairness among all models, they were all trained for 100 epochs using a mini-batch size of 64, SGD optimiser and the exact same backbones. Given the nature of the MSLR method [17], different learning rates for each modality were used: 0.001 for audio, 0.01 for visual, and 8e-5 for text. For the rest of the methods, including AMST and its sub-modules, a learning rate of 0.001 was used for all modalities. Moreover, all SOTA methods were evaluated using the exact parameters specified in their original works, and their official code was used whenever available. All modalities' skip parameters require fine-tuning and their optimal values may differ depending on each modality's informativeness and task at hand. The visual skip parameter was set to 1 in all datasets (i.e., we do not skip its training). Concerning audio, its skip parameter was set to 5 for CREMA-D, 2 for AVE, 4 for IEMOCAP and 6 for UR-FUNNY. For text, we set the text skip parameter to 10 for MVSA, IEMOCAP and UR-FUNNY. Moreover, all modality skip parameters were the same for both the alternating and joint skip methods. Regarding the final prediction, we followed a standard even split approach for all methods that adopted late fusion prediction (MLA and AMST).

Experiments were conducted to compare the performance of AMST, state-of-the-art methods, and their unimodal counterparts. We further make comparisons of the performances of AMST, by excluding its key components, i.e., **(1)** Baseline Joint Module, **(2)** Skip Training in Joint Module (AMST-Joint), **(3)** Baseline Alternating Module, **(4)** Skip Training in Alternating Module (AMST-Alt), **(5)** Alternating and Joint Modules Without Skip Training and **(6)** Skip Training in the Complete Architecture (AMST-Full). All experiments were performed on an A100 GPU and the average training time per epoch is shown in Table 3. Finally, in Table 1 and Table 3, we include a baseline Naïve method, which uses a joint sum fusion approach without any optimisation, where modality embeddings are summed before being used in the classification head. This baseline serves as a reference point for performance comparisons across all datasets.

Table 1: Comparison of accuracies of unimodal models, state-of-the-art methods, and AMST architectures on the used test sets. Results are reported as the average over 3 random seeds, with standard error. The best results for each dataset are displayed in bold, and the second-best are underlined.

Method	CREMA-D	AVE	MVSA	IEMOCAP	URFUNNY
Audio	59.1 ±0.49	55.3 ± 0.56	-	49.2 ± 1.41	59.7 ± 0.64
Visual	60.5 ±0.75	28.5 ± 0.64	57.7 ± 1.21	46.5 ± 0.53	49.2 ± 0.78
Text	-	-	70.8 ± 0.10	62.3 ± 0.30	68.9 ± 0.05
Naïve	62.9 ± 1.24	55.9 ± 0.46	71.1 ± 0.10	67.3 ± 1.20	70.0 ± 0.26
OGM-GE	65.5± 1.46	56.1 ± 1.24	<u>72.7</u> ± 0.55	67.5 ± 0.44	70.3 ± 0.38
PMR	67.5 ± 1.65	58.8 ± 1.55	**73.6** ± 0.20	-	-
MSLR	71.6 ± 0.82	61.1 ± 0.70	71.0 ± 0.10	61.1 ± 0.70	69.0 ± 0.58
MLA	74.8 ± 0.30	61.9 ± 0.44	71.8 ± 0.10	63.9 ± 0.62	65.6 ± 0.24
AMST-Alt	<u>79.0</u> ± 0.55	62.1 ± 0.76	72.5 ± 0.31	67.5 ± 0.20	70.8 ± 0.34
AMST-Joint	78.2 ± 0.03	<u>63.2</u> ± 0.46	72.5 ± 0.79	<u>67.6</u> ± 0.99	**72.1** ± 0.46
AMST-Full	**80.5** ± 0.77	**65.8** ± 0.38	72.6 ± 0.43	**68.4** ± 0.54	<u>71.6</u> ± 0.13

5 Results and Discussion

5.1 Comparison with State-of-the-Art Methods

Table 1 shows the performance of AMST against state-of-the-art methods and unimodal models. We separate the AMST architecture into three components: **AMST-Alt**, Figure 1 (a), corresponds to the Alternating Module; **AMST-Joint**, Figure 1 (b), corresponds to the Joint Module; **AMST-Full**, Figure 1 (c), consists of using both modules in a complementary fashion. For clearer comparisons, it is important to note that only MLA and AMST-Alt utilise an alternating architecture with separately optimised modality-specific losses. In contrast, AMST-Joint, MSLR, PMR, OGM-GE, and the Naïve methods follow a joint training approach, relying on a single multimodal loss. Moreover, PMR does not provide an implementation for more than two modalities, which explains why its results are missing for IEMOCAP and UR-FUNNY.

As evident from the results, all methods outperform their unimodal counterparts, except for MSLR in IEMOCAP and UR-FUNNY, and MLA in UR-FUNNY. AMST's alternating and joint architectures consistently surpass their respective alternatives under the same conditions across all datasets, except for MVSA-Single. In CREMA-D, AVE, and IEMOCAP, the full AMST architecture achieves the best performance, with both the alternating (AMST-Alt) and joint (AMST-Joint) variants performing similarly while still outperforming other methods. For the UR-FUNNY dataset, AMST-Joint emerges as the best-performing architecture, with AMST-Alt and AMST-Full following closely behind.

The primary reason AMST architectures do not outperform other methods on the MVSA-Single dataset is that the visual modality is less informative than the textual one. A possible explanation for this discrepancy in modality

Table 2: Average accuracies over 3 random seeds with standard error for AMST, with and without its main components, across datasets. CREMA-D and AVE use audio-visual modalities, MVSA uses visual-text, while IEMOCAP and UR-FUNNY incorporate all three. In the "Alt." (alternating), "Joint," and "Skip" columns, a tick indicates inclusion of the respective method. For each dataset, the best result is in bold and the second-best is underlined.

Exp	Alt.	Joint	Skip	CREMA-D	AVE	MVSA	IEMOCAP	UR-FUNNY
1	-	✓	-	64.0 ± 1.45	59.6 ± 1.76	71.7 ± 0.90	65.3 ± 0.35	68.3 ± 0.25
2	-	✓	✓	78.2 ± 0.03	63.2 ± 0.46	<u>72.5</u> ± 0.79	67.6 ± 0.99	**72.1** ± 0.46
3	✓	-	-	74.8 ± 0.35	61.9 ± 1.13	71.0 ± 0.93	65.5 ± 1.69	66.1 ± 0.53
4	✓	-	✓	<u>79.0</u> ± 0.55	62.1 ± 0.76	<u>72.5</u> ± 0.31	67.5 ± 0.20	70.8 ± 0.34
5	✓	✓	-	75.9 ± 1.39	<u>64.3</u> ± 0.58	**72.6** ± 0.07	<u>67.7</u> ± 1.27	68.4 ± 0.21
6	✓	✓	✓	**80.5** ± 0.77	**65.8** ± 0.38	**72.6** ± 0.43	**68.4** ± 0.54	<u>71.6</u> ± 0.13

informativeness, compared to other datasets, is that in MVSA, the visual modality consists of a single image per sample, rather than multiple frames of a video. Additionally, training the visual modality is a significantly prolonged process, requiring substantial time to reach convergence. Consequently, models that favour a bias toward the faster learning modality may achieve superior results as the inherent imbalance, in this case, tends to be advantageous. Although AMST is not the best-performing model for MVSA, all methods exhibit similar performance on this dataset, approaching the baseline unimodal text accuracy.

In contrast, AMST-Alt demonstrates a clear advantage over the standard MLA alternating architecture. In a three-modality scenario, as seen in IEMOCAP and UR-FUNNY, the MLA method underperforms compared to other approaches, highlighting its limitations in handling more complex multimodal interactions. With the introduction of skip training and the mitigation of bias toward faster-learning modalities, AMST-Alt demonstrates that the alternating architecture can achieve performance comparable to joint models on these datasets.

In conclusion, AMST-Alt outperforms MLA in every scenario, highlighting the effectiveness of skip training in alternating and joint training approaches.

5.2 Ablation Study

Table 2 shows results from including and excluding key components of the full AMST architecture. These experiments used audio, visual, and text modalities.

Rows 1 & 2 of Table 2 present the results of the joint method without and with skip training, respectively. Across all datasets, we observe a significant improvement in performance with the incorporation of skip training. Notably, in the CREMA-D dataset, the model's accuracy increases by 14.2 percentage points. This substantial improvement highlights the effectiveness of balancing modalities based on convergence rates in joint architectures, particularly in scenarios where modalities provide comparable levels of informativeness. Prior to introducing

Table 3: Comparisons on average training time per epoch (in seconds, on GPU, averaged over 100 epochs) between state-of-the-art methods and AMST architectures. The percentage of performance improvements or degradations between AMST architectures and the Naïve baseline are reported with (+x%) or (-x%) respectively in the last three rows.

Method	CREMA-D	AVE	MVSA	IEMO-CAP	UR-FUNNY
Naïve	20.3	14.2	27.3	36.9	53.0
MSLR	20.3	14.2	27.3	37.7	53.3
OGM	20.9	15.0	28.0	37.8	54.5
PMR	41.0	65.0	34.2	-	-
MLA	20.8	16.5	27.4	37.6	52.8
AMST-Alt	15.4(+**24%**)	12.4(+**13%**)	16.9(+**38%**)	26.1(+**29%**)	34.1(+**36%**)
AMST-Joint	14.7(+**28%**)	12.2(+**14%**)	17.0(+**38%**)	25.4(+**31%**)	33.9(+**36%**)
AMST-Full	25.1(-24%)	17.1(-20%)	32.0(-17%)	45.7(-24%)	63.6(-20%)

skip training, the underutilised visual modality was unable to fully contribute its informative potential to the final prediction.

A notable improvement is also observed in the alternating architecture (rows 3 & 4 of Table 2, without and with skip training, respectively). The alternating approach facilitates the learning of all modalities by optimising separate modality-specific losses. However, by further mitigating the imbalance caused by varying convergence rates, skip training proves advantageous over scenarios without it.

With the inclusion of skip training, the Alternating and Joint architectures achieve similar performance (rows 2 & 4), closing the gap where joint architectures previously underperformed, especially in the CREMA-D dataset (rows 1 & 3). This further highlights the benefits of skip training.

Finally, rows 5 & 6 of Table 2 present the results of integrating Alternating and Joint Modules, without and with skip training, respectively. Row 5 highlights the advantages of combining these architectures, consistently outperforming the standalone Alternating and Joint cases (rows 1 & 3). Furthermore, as expected, the inclusion of skip training in both modules enhances performance, yielding the best overall results across all datasets, except for the UR-FUNNY dataset.

5.3 Computational Expenses

Table 3 presents the training times for every experiment. We can observe that, compared to the baseline (the Naïve method), most approaches introduce minimal overhead, except for the PMR method, which nearly doubles the training time due to the additional cost of computing prototypes. Notably, our proposed methods, both AMST-Alt and AMST-Joint, significantly reduce the average training time with improvements ranging from around **13%** to **38%** (AMST-Alt and AMST-Joint rows of Table 3) compared to the fastest method (Naïve), while achieving superior results. Moreover, our **Full** architecture (AMST-Full), despite

incorporating **4 encoders** (2x as many as other methods), only increases the training time by approximately **17-25%** (AMST-Full row of Table 3) compared to the fastest method (Naïve), demonstrating skip training's efficiency.

6 Conclusion

In this work, we proposed a novel method called AMST to address one aspect of modality imbalance, indicated by different convergence rates across modalities. AMST optimises the multimodal learning process without compromising modality-specific information by having (1) independent training for each of the modalities in the Alternating Module, (2) conventional joint training in the Joint Module, and (3) skip training in both Alternating and Joint Modules.

Our experiments demonstrated that skip training in AMST decreases the dominance gap between modalities in terms of convergence rate, reduces the overall training cost, and provides the complementary benefits of alternating and conventional multimodal joint training practices. Furthermore, we demonstrate AMST's efficacy in scenarios with up to three modalities.

As a consequence of the implementation of skip training in the learning process, an additional hyperparameter, the skip parameter, is created per modality. Currently, fine-tuning each skip parameter is necessary. A future direction is to quantify the convergence rate disparities between modalities and automate the skipping process during training by dynamically adjusting each modality's skip parameter. Finally, it is important to note that improper skip parameter settings may result in suboptimal outcomes, as excessive skipping can lead to under-optimisation of the modalities.

References

1. Busso, C., Bulut, M., Lee, C.C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J.N., Lee, S., Narayanan, S.S.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
2. Cao, H., Cooper, D.G., Keutmann, M.K., Gur, R.C., Nenkova, A., Verma, R.: Crema-d: Crowd-sourced emotional multimodal actors dataset. IEEE Transactions on Affective Computing **5**(4), 377–390 (2014). https://doi.org/10.1109/TAFFC.2014.2336244
3. Fan, Y., Xu, W., Wang, H., Wang, J., Guo, S.: Pmr: Prototypical modal rebalance for multimodal learning. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20029–20038 (2023). https://doi.org/10.1109/CVPR52729.2023.01918
4. Hasan, M.K., Rahman, W., Zadeh, A., Zhong, J., Tanveer, M.I., Morency, L.P., et al.: Ur-funny: A multimodal language dataset for understanding humor. arXiv preprint arXiv:1904.06618 (2019)
5. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

6. Li, H., Li, X., Hu, P., Lei, Y., Li, C., Zhou, Y.: Boosting multi-modal model performance with adaptive gradient modulation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22157–22167 (2023). https://doi.org/10.1109/ICCV51070.2023.02030
7. Lin, X., Wang, S., Cai, R., Liu, Y., Fu, Y., Tang, W., Yu, Z., Kot, A.: Suppress and rebalance: Towards generalized multi-modal face anti-spoofing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 211–221 (June 2024)
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
9. McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th Python in Science Conference,Proceedings of the Python in Science Conference (Dec 2014). https://doi.org/10.25080/majora-7b98e3ed-003
10. Niu, T., Zhu, S., Pang, L., El Saddik, A.: Sentiment analysis on multi-view social data. In: MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22. pp. 15–27. Springer (2016)
11. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced Multimodal Learning via On-the-fly Gradient Modulation . In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8228–8237. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2022). https://doi.org/10.1109/CVPR52688.2022.00806
12. Sun, Y., Mai, S., Hu, H.: Learning to balance the learning rates between various modalities via adaptive tracking factor. IEEE Signal Processing Letters **28**, 1650–1654 (2021). https://doi.org/10.1109/LSP.2021.3101421
13. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: Proceedings of the European Conference on Computer Vision (ECCV) (September 2018)
14. Wang, W., Tran, D., Feiszli, M.: What makes training multi-modal classification networks hard? In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12692–12702 (2020). https://doi.org/10.1109/CVPR42600.2020.01271
15. Wu, N., Jastrzebski, S., Cho, K., Geras, K.J.: Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 24043–24055. PMLR (17–23 Jul 2022)
16. Xu, N., Mao, W.: Multisentinet: A deep semantic network for multimodal sentiment analysis. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. pp. 2399–2402 (2017)
17. Yao, Y., Mihalcea, R.: Modality-specific learning rates for effective multimodal additive late-fusion. In: Muresan, S., Nakov, P., Villavicencio, A. (eds.) Findings of the Association for Computational Linguistics: ACL 2022. pp. 1824–1834. Association for Computational Linguistics, Dublin, Ireland (May 2022). https://doi.org/10.18653/v1/2022.findings-acl.143
18. Zhang, X., Yoon, J., Bansal, M., Yao, H.: Multimodal Representation Learning by Alternating Unimodal Adaptation . In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27446–27456. IEEE Computer Society, Los Alamitos, CA, USA (Jun 2024). https://doi.org/10.1109/CVPR52733.2024.02592