Adaptive Unimodal Regulation
for Balanced Multimodal Information Acquisition

Chengxiang Huang[1,†]   Yake Wei[2,†]   Zequn Yang[2]   Di Hu[2,*]

[1]Beijing University of Posts and Telecommunications   [2]Renmin University of China

huangchengxiang2021@bupt.edu.cn, {yakewei,zqyang,dihu}@ruc.edu.cn

Abstract

Sensory training during the early ages is vital for human development. Inspired by this cognitive phenomenon, we observe that the early training stage is also important for the multimodal learning process, where dataset information is rapidly acquired. We refer to this stage as the prime learning window. However, based on our observation, this prime learning window in multimodal learning is often dominated by information-sufficient modalities, which in turn suppresses the information acquisition of information-insufficient modalities. To address this issue, we propose **Info***rmation Acquisition* **Reg***ulation (InfoReg), a method designed to balance information acquisition among modalities. Specifically, InfoReg slows down the information acquisition process of information-sufficient modalities during the prime learning window, which could promote information acquisition of information-insufficient modalities. This regulation enables a more balanced learning process and improves the overall performance of the multimodal network. Experiments show that InfoReg outperforms related multimodal imbalanced methods across various datasets, achieving superior model performance. The code is available at* https://github.com/GeWu-Lab/InfoReg_CVPR2025.

1. Introduction

Learning during early developmental ages in humans and animals is important for skill impairments [16, 22, 44]. Similarly, in deep learning, recent studies have found that models learn in stages, with early learning being especially important for effective information acquisition [2, 19].

The above circumstance motivates us to investigate the process of information acquisition in the multimodal scenario. Due to the presence of multiple modalities, multimodal learning networks are expected to capture sufficient information from each modality [12, 40]. To observe the
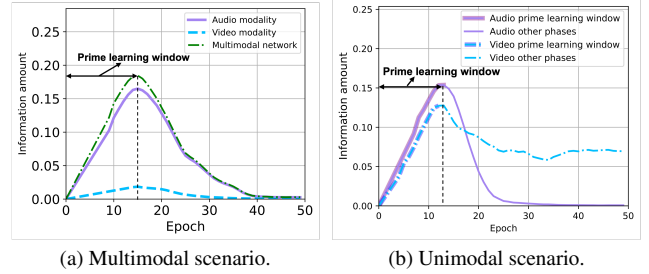

(a) Multimodal scenario.   (b) Unimodal scenario.

Figure 1. **(a).** Information amount variation of the audio encoder, video encoder, and multimodal model during the training process on CREMA-D [7]. **(b).** Information amount variation of the audio and video modalities when trained independently on CREMA-D.

information acquisition of different modalities, we use the trace of the Fisher Information Matrix [2, 10] to investigate the information amount of each modality in multimodal models. However, based on our observation, the amount of information in multimodal models is not well consistent with intuitive expectations. *Firstly*, as shown in Figure 1a, the green curve, representing the overall multimodal network, demonstrates a rapid increase in information amount during the early periods. Information acquisition is most rapid in this early stage, which we refer to as *the prime learning window*, reaching a peak at the end of this period, followed by a subsequent decline. *Secondly*, for audio modality, its overall trend closely aligns with the overall multimodal model, and shows a high information acquisition amount during the prime learning window. These findings align with our expectation that modality information could be acquired effectively within the prime learning window. However, the video modality, represented by the blue curve, shows a much lower information acquisition amount during the prime learning window. Although the video modality is capable of effective information acquisition in the unimodal scenario in Figure 1b, it fails to do so in the multimodal scenario when trained jointly with the audio modality. These observations suggest that information-insufficient modalities, like video, experience suppressed information acquisition during the prime learn-

---

[†]Equal contribution. [*]Corresponding author.

Method	Accuracy
Video in Joint training (50 epochs)	35.14
Video in Joint training (100 epochs)	35.36
Video in InfoReg (50 epochs)	**49.65**

Table 1. Comparison of performance for the video modality between Joint training and InfoReg on the CREMA-D dataset. Simply extending the training time (100 epochs) cannot compensate for the suppressed video modality.

ing window due to the stronger information acquisition capacity of information-sufficient modalities, such as audio. Moreover, as the multimodal model's capacity for information acquisition diminishes in later stages, the imbalance observed during the prime learning window cannot be compensated for by simply extending the training time, as demonstrated in Table 1.

Recent studies have investigated the problem of imbalanced learning across modalities in multimodal learning [18, 32, 41]. Although several studies have made progress in addressing this issue [8, 28, 35, 42], their methods typically apply adjustments across the entire training process without recognizing the importance of the prime learning window. Consequently, the effectiveness of these methods is significantly constrained. Our research reveals that the prime learning window plays a vital role in multimodal learning. In this window, there is a significant imbalance in information acquisition across modalities. Therefore, effective adjustment within this window is essential for achieving balanced information acquisition across modalities.

Based on the above analysis, information-insufficient modalities experience a significantly reduced information acquisition amount during the prime learning window due to suppression from information-sufficient modalities. To address this imbalance, we slow the information acquisition rate of information-sufficient modalities within this window, thereby allowing information-insufficient modalities to improve their acquisition. Hence, we propose our method, **Info**rmation Acquisition **Reg**ulation (InfoReg). In InfoReg, the process begins by determining whether information-sufficient modalities are within the prime learning window. If so, a unimodal regulation term is applied to regulate the Fisher Information [2, 10], thereby restricting these modalities acquire information. As a result, InfoReg promotes a more balanced information acquisition across modalities, enhancing the overall performance of the model. Our extensive experiments across multiple datasets show that InfoReg achieves superior performance and improves modality balance by helping information-insufficient modalities acquire more information during the prime learning window. Additionally, InfoReg enhances feature quality, supporting prior research on the link between early-stage information acquisition and robust feature representation [1, 2]. These results collectively validate

the effectiveness of our approach.

Our main contributions are summarized as follows:

- **Firstly**, we identify the prime learning window in multimodal learning, a critical period where imbalances in information acquisition significantly impact modality balance and overall performance.
- **Secondly**, We analyze the imbalance of Fisher Information among modalities and propose InfoReg, a method that regulates the information acquisition of information-sufficient modalities during the prime learning window.
- **Finally**, we validate InfoReg on multiple datasets and settings, demonstrating its considerable improvement while maintaining balanced performance across modalities.

2. Related Work

2.1. Multimodal imbalance learning

Jointly training multiple modalities is intuitively expected to enhance performance [30]. However, imbalanced learning across modalities poses substantial challenges, hindering the effective training of multimodal networks [3, 17, 36, 37]. Previous studies have investigated the imbalanced learning problem across modalities in multimodal learning networks [32, 41], where certain modalities tend to dominate, limiting the learning of other modalities. This imbalance can degrade overall performance and lead to significant disparities between modalities, sometimes even causing multimodal learning to underperform compared to single-modality scenarios [32]. To address this problem, many methods [28, 32, 41–43] have been proposed, primarily focusing on balancing the optimization of each modality throughout the training process. BalanceBench [47] further categorizes these methods based on their distinct characteristics, offering a comprehensive framework to evaluate their effectiveness and limitations. Specifically, OGM [32] balances modalities by adjusting the gradients of well-learned modalities. MMPareto [42] leverages an optimized Pareto front to balance the performance across modalities, aiming to improve the generalization of multimodal models. Despite the successes of previous work, they have largely overlooked the impact of different training periods on information acquisition across various modalities. We have recognized this issue and designed a method accordingly. By slowing down the information acquisition rate of information-sufficient modalities, our approach alleviates the suppression of information-insufficient modalities, allowing them to acquire more information in the early stages, effectively balancing the performance of different modalities. This leads to considerable improvements in both overall performance and modality balance.

2.2. Early stages in deep learning

The learning process of deep learning models consists of two main phases: an initial phase of information acquisition from the dataset, followed by a phase of gradual information compression or forgetting [1, 19, 21, 48]. Recent studies highlight the crucial role of early-stage learning in shaping feature representation and overall model performance [11, 13, 20, 27, 49], similar to findings in neuroscience [6, 16, 22, 45]. Further, the information missed during the early stages cannot be recovered by extending the training duration later on [2]. We define this early important learning period as the prime learning window. However, previous research has overlooked the importance of the prime learning window in multimodal learning. Our findings reveal that balancing information acquisition among modalities during this window leads to considerable performance improvements in multimodal models.

3. Method

3.1. Multimodal learning framework

For convenience, let the dataset be denoted by $D = \{(x_n, y_n)\}_{n=0,1,\ldots,N-1}$, where each x_n contains inputs from M modalities: $x_n = (x_n^1, x_n^2, \ldots, x_n^M)$. The target label $y_n \in \{1, 2, \ldots\}$ represents the class of sample x_n. For each modality m, where $m \in \{1, 2, \ldots, M\}$, the input is processed through the corresponding encoder $\varphi^m(w_m, \cdot)$. Here, w_m are the weights of encoder m. After feature extraction, the outputs are concatenated and passed to a single-layer linear classifier. Finally, one joint multimodal cross-entropy loss $\mathcal{L}_{joint}(w)$ is utilized to optimize the model.

3.2. Fisher Information in multimodal learning

Following previous work [1, 34] that uses the Kullback-Leibler (KL) divergence to measure the information contained in the weights of a unimodal network, the information acquisition process for a single modality m can be evaluated by this metric. Given the posterior distribution $p_{w_m}(y|x; D)$, encoded by the unimodal encoder with weights w_m and its prior distribution $q_{w_m}(y|x)$, the mutual information is defined as follows:

$$D_{KL}(p_{w_m} \parallel q_{w_m}) = \int p_{w_m}(y|x; D) log \frac{p_{w_m}(y|x; D)}{q_{w_m}(y|x)} dy. \tag{1}$$

However, quantifying the information a unimodal encoder acquires from the dataset is challenging because the ground truth prior distribution $q_{w_m}(y|x)$ is not accessible. As an alternative, inspired by [2], the rate of information acquisition from the dataset can be estimated by calculating the KL divergence between distributions encoded by weights at successive moments in training. Specifically, given a perturbation $w_m' = w_m + \delta w_m$. The discrepancy between the

distributions $p_{w_m}(y|x; D)$ and $p_{w_m'}(y|x; D)$ reflects the rate of information acquisition and can be defined as:

$$D_{KL}(p_{w_m} \parallel p_{w_m'}) = \int p_{w_m}(y|x; D) log \frac{p_{w_m}(y|x; D)}{p_{w_m'}(y|x; D)} dy. \tag{2}$$

Further, the KL divergence can be approximated to second order by applying the Taylor expansion:

$$D_{KL}(p_{w_m} \parallel p_{w_m}') \approx \frac{1}{2} \delta w_m^T F_m \delta w_m, \tag{3}$$

where F_m is the Fisher Information Matrix (FIM) [10], and is defined as:

$$F_m = \mathbb{E}_{y \sim p_w} \left[\nabla_{w_m} log p_{w_m}(y|x) \nabla_{w_m} log p_{w_m}(y|x)^T \right]. \tag{4}$$

The FIM plays a crucial role in quantifying the amount of information captured by the deep neural network [2, 26] and acts as a local measure, assessing how small perturbations in the model's parameters influence its output [4]. Additionally, the FIM is a semi-definite approximation of the Hessian matrix, providing insights into the curvature of the loss landscape at a given point during training [31, 38].

In multimodal learning, for a unimodal encoder φ^m, the gradient of the encoder can be expressed as :

$$g_{\varphi^m}(w_m, x_n^m) = \nabla_{w_m} log p_{w_m}(y_n|x_n^m) = \nabla_{w_m} \mathcal{L}_{joint}(w_m). \tag{5}$$

Based on this, the Fisher Information Matrix F_m can be reformulated as:

$$F_m = \mathbb{E}_{x_n^m \sim X^m} \left[g_{\varphi^m}(w_m, x_n^m) g_{\varphi^m}(w_m, x_n^m)^T \right]. \tag{6}$$

However, computing F_m directly is computationally expensive. To address this, we use the trace of the Fisher Information Matrix, denoted as $Tr(F_m)$, to measure the amount of information captured by the deep neural network. This trace can be computed more efficiently and is defined as:

$$Tr(F_m) = \mathbb{E}_{x_n^m \sim X^m} \left[\parallel g_{\varphi^m}(w_m, x_n^m) \parallel^2 \right]. \tag{7}$$

As shown in Figure 1, $Tr(F_m)$ could effectively measure the amount of information acquired and identify the prime learning window.

As illustrated in Figure 3a, information-sufficient modalities will exhibit significantly larger values of g_{φ^m} during the prime learning window. Due to the squared term in Equation 7, these substantial differences in g_{φ^m} between modalities are further amplified, thereby making the imbalance in Fisher Information even more pronounced. This indicates that information-sufficient modalities have a clear advantage in the information acquisition during the prime learning window and dominate the overall information acquisition of the multimodal model.
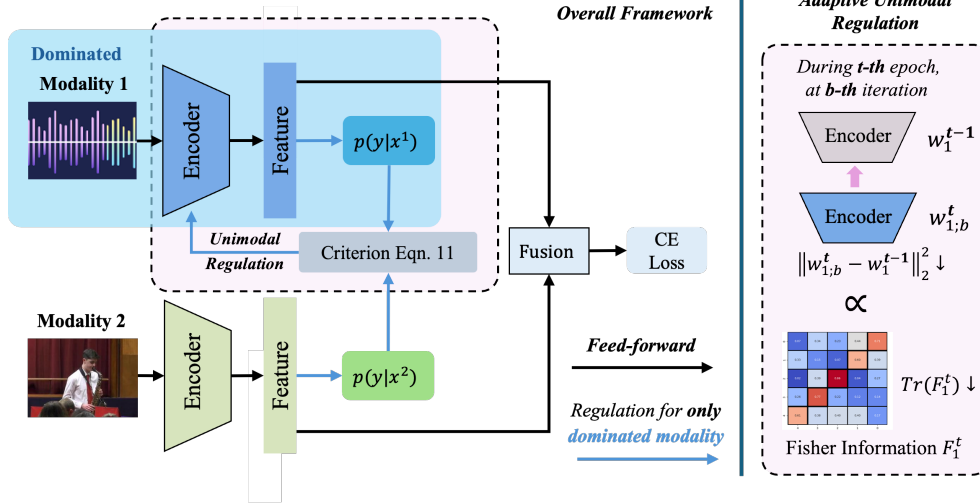
Figure 2. **Overview of InfoReg.** This figure shows the main components and workflow of InfoReg. The left side presents our overall framework, while the right side highlights the adaptive unimodal regulation. During the training, InfoReg first identifies the information-sufficient modalities, then evaluates whether they are in the prime learning window, and finally applies adaptive unimodal regulation.
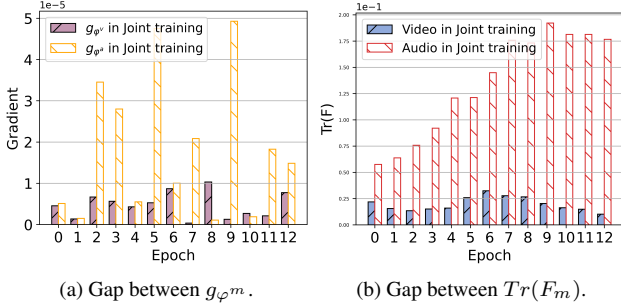

Figure 3. **(a).** The gradient gap between the audio modality and video modality on CREMA-D. **(b).** The $Tr(F_m)$ gap between the audio modality and video modality on CREMA-D.

3.3. Information acquisition regulation

To address the Fisher Information imbalance observed during the prime learning window, our method regulates the Fisher Information of information-sufficient modalities in this important period to slow down their information acquisition, thereby promoting information acquisition in other modalities. Our method consists of the following main components:

- *Evaluate the prime learning window for information-sufficient modalities*: We first identify the information-sufficient modalities and then evaluate whether they fall within the prime learning window.
- *Adaptive unimodal regulation*: For information-sufficient modalities, we apply adaptive unimodal regulation to approximately regulate the Fisher Information.

3.3.1 Evaluating the prime learning window.

Inspired by OGM [32], performance scores are used to identify information-sufficient modalities. Afterward, an as-

sessment is made to determine whether these information-sufficient modalities are in the prime learning window. For each iteration, assume that training has reached epoch t and is currently processing batch b, where $b \in \{0, 1, \dots, B-1\}$ and B represent the total number of batches. The performance score for each modality is given by:

$$s_{m;b}^t = \mathbb{E}_{x_n^m \sim X^m} \left[-log \left(softmax(\varphi^m (x_n^m))_{y_n} \right) \right]. \quad (8)$$

Then, we define its performance gap Δ_m relative to other modalities to determine information-sufficient modalities during the prime learning window. Let C_m represent the number of modalities with performance scores less than $s_{m;b}^t$:

$$C_m = \left| \left\{ m' \in [M] \setminus \{m\} \, ; s_{m';b}^t < s_{m;b}^t \right\} \right|. \quad (9)$$

The performance gap Δ_m can then be expressed as:

$$\Delta_m = \frac{1}{C_m} \sum_{m' \in [M] \setminus \{m\}; s_{m';b}^t < s_{m;b}^t} \left(s_{m;b}^t - s_{m';b}^t \right), \quad (10)$$

where Δ_m measures the average performance difference between m and all lower or equally performing modalities. This ensures that for the lowest-performing modality, where $s_{m;b}^t$ is minimal, Δ_m will be 0.

After identifying information-sufficient modalities based on $s_{m;b}^t$, the following criterion is used to determine whether these modalities are in the prime learning window:

$$\frac{Tr(F_m^{t-1}) - Tr(F_m^{t-2})}{Tr(F_m^{t-1})} > K, \quad (11)$$

where K is a positive hyperparameter that controls the threshold for inclusion. Equation 11 reflects the relative
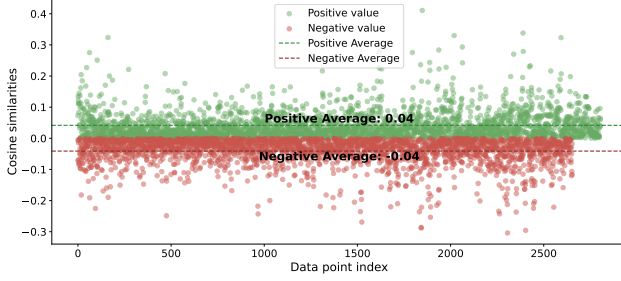
Figure 4. The cosine similarities of gradients across different batches within the prime learning window.

Algorithm 1 Pipeline of InfoReg

Input: Training dataset D, number of epochs T, number of batches B of each batch, hyperparameter β, K
1: **for** $t = 0, 1, \cdots, T-1$ **do**
2:     **if** $t < 2$ **then**
3:         Update model parameters;
4:         Calculate $Tr(F_m^t)$ by Equation 7;
5:         **Continue**;
6:     **end if**
7:     Calculate $\frac{Tr(F_m^{t-1}) - Tr(F_m^{t-2})}{Tr(F_m^{t-1})}$;
8:     **for** $b = 0, 1, \cdots, B-1$ **do**
9:         Randomly selects a batch of data from D;
10:        Calculate the performance scores $s_{m;b}^t$ for different modalities by Equation 8;
11:        Calculate Δ_m by Equation 10;
12:        Decide information-sufficient modalities by $s_{m;b}^t$;
13:        Calculate α by Equation 16;
14:        **if** $\frac{Tr(F_m^{t-1}) - Tr(F_m^{t-2})}{Tr(F_m^{t-1})} > K$ **and** $\Delta_m > 0$ **then**
15:           Calculate the regulation term by Equation 12;
16:           Add adaptive unimodal regulation term;
17:        **end if**
18:        Update model parameters;
19:     **end for**
20:     Calculate $Tr(F_m^t)$ by Equation 7.
21: **end for**

---

changing rate of $Tr(F_m)$. A large value of this rate indicates that the information amount in the unimodal encoder is rapidly increasing. Conversely, a small or negative value suggests that information acquisition has slowed down or is even decreasing.

3.3.2 Adaptive unimodal regulation.

During the prime learning window, information-sufficient modalities dominate the model's information acquisition, limiting the ability of information-insufficient modalities to acquire information. Therefore, it becomes essential to regulate information acquisition of information-sufficient

modalities in this window. However, directly calculating $Tr(F_m^t)$ requires complete gradient information across the entire dataset, making it impractical to regulate effectively. Additionally, calculating $Tr(F_m^t)$ over a full epoch is ineffective for adjusting the model at each iteration. To address this, we introduce a regulation term $P_{m;b}^t$ to approximately regulate the Fisher Information:

$$P_{m;b}^t = \frac{\alpha}{2} \parallel w_{m;b}^t - w_m^{t-1} \parallel^2, \quad (12)$$

where α is a parameter that controls the degree of regulation. This regulation term is proportional to $Tr(F_{m;b}^t)$ for each batch. Let the gradient g_{φ^m} be denoted as g, $Tr(F_{m;b}^t)$ in each batch can be defined as:

$$Tr(F_{m;b}^t) = \frac{1}{b} \sum_{k=0}^{b} \parallel g_k^t \parallel^2 . \quad (13)$$

The regulation term $P_{m;b}^t$ can be written as:

$$
\begin{aligned}
P_{m;b}^t &= \frac{\alpha}{2} \parallel w_{m;b}^t - w_m^{t-1} \parallel^2 \\
&= \frac{\alpha}{2} \parallel -\eta \sum_{k=0}^{b} g_k^t \parallel^2 \\
&= \frac{\alpha \eta^2}{2} \left(\sum_{k=0}^{b} \parallel g_k^t \parallel^2 + 2 \sum_{0 \le z < k \le b} g_z^t (g_k^t)^T \right), \quad (14)
\end{aligned}
$$

where η represents the learning rate. The high dimensionality of g_b^t results in the gradients $g_z^t (g_k^t)^T$ becoming approximately orthogonal for any two batches z and k. The detailed proof is provided in the Appendix A. As illustrated in Figure 4, we compare the cosine similarities between gradients of different batches, and the results confirm that they are approximately orthogonal. Then, $P_{m;b}^t$ can be approximated as:

$$P_{m;b}^t = \frac{\alpha \eta^2}{2} \sum_{k=0}^{b} \parallel g_k^t \parallel^2 . \quad (15)$$

According to the Equation 13 and Equation 15, the unimodal regulation term $P_{m;b}^t$ regulates $Tr(F_{m;b}^t)$, thereby limiting the amount of information acquired by the information-sufficient modalities. To modulate the impact of $P_{m;b}^t$, we define α as a dynamic parameter:

$$\alpha = exp\left(\beta * \tanh\left(\Delta_m\right)\right), \quad (16)$$

where β is a hyperparameter controlling the sensitivity of α to the performance gap Δ_m. A higher value of α results in stronger regulation, thereby tightly constraining the information acquisition for information-sufficient modalities. This adaptive regulation ensures balance and prevents any single modality from dominating during the prime learning window. Additional analysis of the regulation term is provided in Appendix B. Overall, our method is shown in Algorithm 1 and illustrated in Figure 2.

Method	CREMA-D			Kinetics Sounds		
	Accuracy	Acc audio	Acc video	Accuracy	Acc audio	Acc video
Joint training	66.61	58.99	35.14	65.67	53.13	36.01
OGM [32]	68.70	56.84	39.52	<u>66.63</u>	53.39	<u>40.16</u>
Greedy [46]	67.82	59.17	40.17	66.54	53.15	37.82
PMR [9]	66.92	57.83	38.91	66.33	<u>53.42</u>	36.17
AGM [28]	<u>69.71</u>	<u>59.32</u>	<u>43.72</u>	66.54	53.12	37.24
InfoReg	**71.90**	**60.03**	**49.65**	**69.31**	**54.16**	**44.73**

Table 2. **Comparison with imbalanced multimodal learning methods.** All the methods only use one multimodal loss. Bold and underline represent the best and second best respectively.

Method	CREMA-D			Kinetics Sounds		
	Accuracy	Acc audio	Acc video	Accuracy	Acc audio	Acc video
Joint training	66.61	58.99	35.14	65.67	54.13	36.01
Joint training*	70.81	60.52	55.23	68.71	55.23	44.18
G-Blending [41]	69.11	60.14	51.29	68.33	54.22	42.31
MMPareto [42]	<u>73.08</u>	<u>60.83</u>	<u>58.92</u>	<u>71.11</u>	<u>56.47</u>	<u>53.39</u>
InfoReg*	**75.71**	**61.63**	**61.22**	**72.03**	**57.21**	**53.57**

Table 3. **Comparison with imbalanced multimodal learning methods with unimodal loss. Joint training*** and **InfoReg*** denote Joint training with unimodal loss and InfoReg with unimodal loss, repectively. Bold and underline represent the best and second best.

4. Experiments

4.1. Dataset and experimental settings

CREMA-D [7] is an emotion recognition dataset with recordings of actors expressing six emotions, providing audio-visual samples to examine how auditory and visual cues convey emotion. **Kinetics Sounds** [5, 23] is designed for human action recognition, featuring 31 action classes from varied video sources, allowing analysis of audio-visual integration in dynamic activity recognition. **CMU-MOSI** [50] is a sentiment analysis dataset with short video clips including audio, visual, and text modalities, suitable for exploring multimodal sentiment expression.

For our model architecture, we employ ResNet-18 [15] as the backbone for the CREMA-D and Kinetics Sounds, while for the CMU-MOSI, we use a transformer-based model [29]. All models are trained from scratch to ensure that the feature extraction processes are fully optimized for our specific tasks and datasets. Additionally, we implement a late fusion method to integrate uni-modal features from different modalities.

4.2. Comparison with related imbalanced methods

To evaluate the effectiveness of InfoReg in addressing information acquisition imbalance across modalities, we compare InfoReg with several imbalanced multimodal learning approaches, including G-Blending [41], OGM [32], Greedy [46], PMR [9], AGM [28], and MMPareto [42]. Of these methods, G-Blending [41] and MMPareto [42] utilize both unimodal and multimodal losses, while OGM [32], Greedy [46], PMR [9], and AGM [28] rely only on unimodal loss. In our evaluation framework, **Joint training** denotes the widely used baseline for imbalanced multimodal learning, utilizing concatenation fusion with a single multimodal cross-entropy loss function [28, 32]. Meanwhile, **Joint training*** represents the scenario where both unimodal and multimodal joint losses are applied simultaneously.

As shown in Table 2 and Figure 5a, we conduct experiments on CREMA-D and Kinetics Sounds using InfoReg and several related methods that employ only multimodal loss. Based on the results, we first observe that all imbalanced methods achieve improved performance, indicating the significance of the modality imbalance issue and the need for balancing unimodal learning. Furthermore, InfoReg consistently outperforms other methods, especially with a notable improvement in the video modality. This suggests that regulating information acquisition during the prime learning window effectively balances information gain across modalities, enabling information-sufficient modalities to acquire more information and enhancing overall model performance.

Additionally, we provide experiments with methods incorporating both multimodal loss and unimodal loss. Here, **Joint training*** and **InfoReg*** indicate both multimodal loss and unimodal loss are applied. As shown in Table 3, all methods with unimodal loss achieve notable improvements over Joint training. This improvement is attributed to the inclusion of unimodal loss, which facilitates more efficient information retrieval from individual modalities, thereby
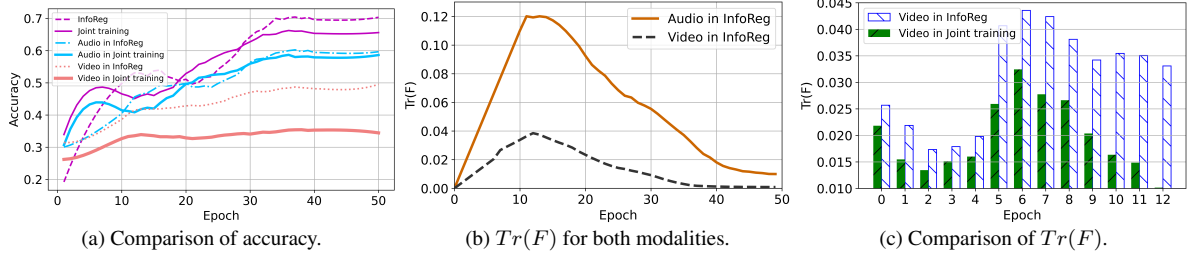
(a) Comparison of accuracy.

(b) $Tr(F)$ for both modalities.

(c) Comparison of $Tr(F)$.

Figure 5. **(a).** The overall accuracy, audio accuracy, and video accuracy of InfoReg are compared with Joint training. **(b).** The value of $Tr(F)$ in InfoReg for both modalities. **(c).** The value of $Tr(F)$ of the video modality in InfoReg compared with that of Joint training. **All experiments are conducted on CREMA-D.**

(a) InfoReg.

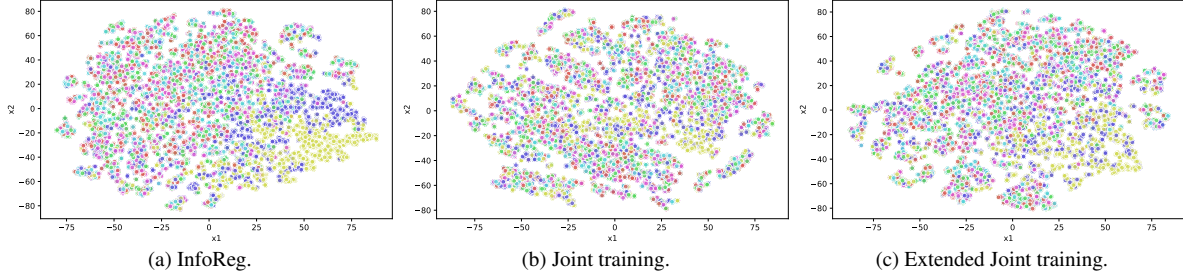(b) Joint training.

(c) Extended Joint training.

Figure 6. The representations of the video modality on CREMA-D by t-SNE [39] in InfoReg and Joint training are shown. "Extended Joint training" refers to Joint training that is extended to 100 epochs. Additional t-SNE representations are provided in Appendix C.

enhancing overall performance. Moreover, InfoReg* continues to outperform other approaches and achieve better modality balance. This indicates that, with unimodal assistance, InfoReg can still effectively balance information acquisition across modalities, thereby achieving good modality balance and model performance.

4.3. Evaluating information acquisition

We conduct experiments on CREMA-D to evaluate the effectiveness of InfoReg in enhancing information acquisition for information-insufficient modalities during the prime learning window. **Firstly**, As shown in Figure 5b, we observe that both modalities exhibit gradually growing values of $Tr(F)$ within the prime learning window, indicating that each modality acquires sufficient information early on. On one hand, the information-sufficient modality, audio, continues to acquire adequate information despite the regulation during the prime learning window, allowing it to maintain good performance throughout training. On the other hand, the information-insufficient modality also acquires sufficient information, as InfoReg alleviates the suppression of its information acquisition by the information-sufficient modality. **Secondly**, we compare the $Tr(F)$ values for the video modality in InfoReg during the prime learning window with those in Joint training. As shown in Figure 5c, the $Tr(F)$ values for the video modality in InfoReg are consistently higher than those in Joint training, indicating a considerable improvement in information acquisition for the video modality during the prime learning window. This enhancement enables the multimodal model to learn more

Method	CREMA-D		
	WTP	**PLW**	**Other periods**
Joint training	66.61	-	-
OGM [32]	68.70	68.76	67.13
AGM [28]	**69.71**	<u>69.54</u>	**67.41**
PMR [9]	66.92	66.99	66.57
InfoReg	<u>69.03</u>	**71.90**	<u>67.22</u>

Table 4. **Comparison of performance.** "WTP" and "PLW" denotes the whole training process and the prime learning window respectively. "Other periods" indicates adjustments made outside the prime learning window. Bold and underline represent the best and second best.

Method	Fusion strategies			
	Gated	**SUM**	**FiLM**	**Concat**
Joint training	65.32	64.38	66.67	66.61
InfoReg	69.18	70.12	70.23	71.90

Table 5. **Comparison of different fusion strategies.**

comprehensive information during the prime learning window, thereby improving the model's overall performance.

4.4. Importance of the prime learning window

Unlike other methods, InfoReg introduces adjustments exclusively during the prime learning window. To evaluate the importance of the prime learning window in addressing the imbalanced multimodal learning problem, we compare the performance of several related methods with adjustments made exclusively during the prime learning window and during other periods. **Firstly**, all methods show consider-

Method	CMU-MOSI	
	Accuracy	Macro F1
Joint training	61.09	60.74
OGM [32]	<u>61.88</u>	<u>61.32</u>
PMR [9]	61.47	60.98
AGM [28]	61.39	60.43
InfoReg	**62.31**	**62.03**

Table 6. **Comparison with imbalanced multimodal learning methods on the CMU-MOSI dataset.** Bold and underline represent the best and second best respectively.
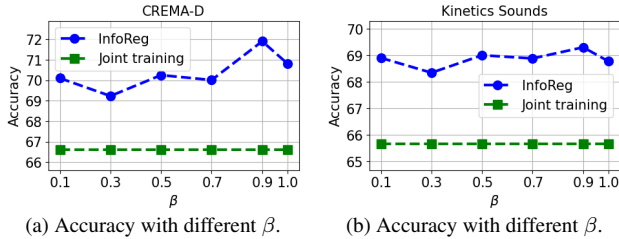

(a) Accuracy with different β.   (b) Accuracy with different β.

Figure 7. **(a).** The overall accuracy of different β on CREMA-D. **(b).** The overall accuracy of different β on Kinetics Sounds.

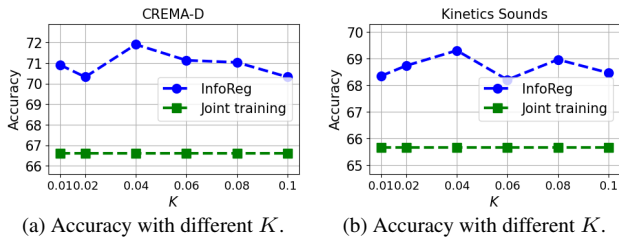(a) Accuracy with different K.   (b) Accuracy with different K.

Figure 8. **(a).** The overall accuracy of different K on CREMA-D. **(b).** The overall accuracy of different K on Kinetics Sounds.

able improvements when adjustments are made exclusively during the prime learning window. Specifically, the performance achieved by these methods during the prime learning window is comparable to that achieved by adjusting throughout the whole training process. This highlights the necessity of making adjustments within the prime learning window. **Secondly**, adjustments made during other periods show significantly lower performance compared to those made during the prime learning window and the whole training process, indicating that late-stage adjustments are unnecessary due to the lack of additional information at this stage. Overall, our method focuses on balancing information acquisition exclusively within the prime learning window, effectively enhancing underrepresented modalities during this period and resulting in strong performance.

To further evaluate the importance of the prime learning window, we use t-SNE [39] to compare the features learned by the video modality using InfoReg on CREMA-D against those learned through Joint training and extended Joint training. As illustrated in Figure 6, InfoReg yields higher-quality features due to sufficient information acquisition in the prime learning window.This aligns with previous studies, which have concluded that the quality of fea-

ture learning is highly correlated with the amount of information acquired early in training [1]. Notably, as shown in Figure 6b and Figure 6c, extending the training period does not compensate for the information loss experienced by the video modality during the prime learning window. Our experiments highlight the critical role of the prime learning window in information acquisition.

4.5. The influence of fusion strategies

We evaluate the performance of InfoReg with four different fusion strategies, including Gated [24], SUM, FiLM [33], and Concat. For the Gated, SUM, and FiLM strategies, we measured the performance of each modality individually by zeroing out the features of other modalities. As shown in Table 5, InfoReg demonstrates consistently robust performance across various fusion strategies. This highlights the adaptability and scalability of InfoReg.

4.6. Extension to more complex settings

To validate that our method continues to perform well in more complex settings, we conducted experiments on CMU-MOSI using a Transformer architecture following [29]. As shown in Table 6, InfoReg continues to perform effectively in these more challenging scenarios, outperforming other related methods. This demonstrates the good scalability of InfoReg, highlighting its ability to extend to more complex transformer-based architectures and scenarios involving more than two modalities. Further experiments exploring dominant modalities and scenarios requiring intermodality cooperation are provided in Appendix D.

4.7. Hyperparameter sensitivity analysis

The hyperparameter β, controlling the regulation term's strength, is selected from $\{0.1, 0.3, 0.5, 0.7, 0.9, 1.0\}$. A larger β increases the regulation degree of the information-sufficient modalities. Our results (Figure 7) show that $\beta = 0.9$ provided the best accuracy. Similarly, the hyperparameter K, which serves as the threshold for determining whether the model is within the prime learning window, is chosen from $\{0.01, 0.02, 0.04, 0.06, 0.08, 0.1\}$. Figure 8 indicates that $K = 0.04$ yields the best performance.

5. Conclusion

In this paper, we identify that there is a prime learning window in multimodal learning, and one modality's information acquisition can be suppressed by others during this stage. Then, we propose the Information Acquisition Regulation algorithm. It aims to address the imbalance in information acquisition across modalities by regulating the acquisition rates of information-sufficient modalities during the prime learning window. Our method promotes a more balanced learning process, accordingly enhancing model performance. Experiments across multiple datasets show that our method alleviates imbalanced multimodal learning and then achieves superior performance.

6. Acknowledgment

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. 2, 3, 8

[2] Alessandro Achille, Matteo Rovere, and Stefano Soatto. Critical learning periods in deep networks. In *International Conference on Learning Representations*, 2018. 1, 2, 3

[3] Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. Scaling laws for generative mixed-modal language models. In *International Conference on Machine Learning*, pages 265–279. PMLR, 2023. 2

[4] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*. American Mathematical Soc., 2000. 3

[5] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE international conference on computer vision*, pages 609–617, 2017. 6

[6] Gustavo Arriaga. *Of Mice, Birds, and Men: The Mouse Ultrasonic Song System and Vocal Behavior*. PhD thesis, Duke University, 2011. 3

[7] 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. 1, 6

[8] Chenzhuang Du, Tingle Li, Yichen Liu, Zixin Wen, Tianyu Hua, Yue Wang, and Hang Zhao. Improving multimodal learning with uni-modal teachers. *arXiv preprint arXiv:2106.11059*, 2021. 2

[9] Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, and Song Guo. Pmr: Prototypical modal rebalance for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20029–20038, 2023. 6, 7, 8

[10] Ronald Aylmer Fisher. Theory of statistical estimation. In *Mathematical proceedings of the Cambridge philosophical society*, pages 700–725. Cambridge University Press, 1925. 1, 2, 3

[11] Jonathan Frankle, David J Schwab, and Ari S Morcos. The early phase of neural network training. *arXiv preprint arXiv:2002.10365*, 2020. 3

[12] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10457–10467, 2020. 1

[13] Aditya Sharad Golatkar, Alessandro Achille, and Stefano Soatto. Time matters in regularizing deep networks: Weight decay and data augmentation affect early learning dynamics, matter little near convergence. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[14] Tal Hassner, Yossi Itcher, and Orit Kliper-Gross. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 1–6. IEEE, 2012. 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[16] Takao K Hensch. Critical period regulation. *Annu. Rev. Neurosci.*, 27(1):549–579, 2004. 1, 3

[17] 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 *International conference on machine learning*, pages 9226–9259. PMLR, 2022. 2

[18] Aya Abdelsalam Ismail, Mahmudul Hasan, and Faisal Ishtiaq. Improving multimodal accuracy through modality pretraining and attention. *arXiv preprint arXiv:2011.06102*, 2020. 2

[19] Stanisław Jastrzębski, Zachary Kenton, Nicolas Ballas, Asja Fischer, Yoshua Bengio, and Amos Storkey. On the relation between the sharpest directions of dnn loss and the sgd step length. *arXiv preprint arXiv:1807.05031*, 2018. 1, 3

[20] Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020. 3

[21] Stanislaw Jastrzebski, Devansh Arpit, Oliver Astrand, Giancarlo B Kerg, Huan Wang, Caiming Xiong, Richard Socher, Kyunghyun Cho, and Krzysztof J Geras. Catastrophic fisher explosion: Early phase fisher matrix impacts generalization. In *International Conference on Machine Learning*, pages 4772–4784. PMLR, 2021. 3

[22] Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven Siegelbaum, A James Hudspeth, Sarah Mack, et al. *Principles of neural science*. McGraw-hill New York, 2000. 1, 3

[23] 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. 6

[24] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 8

[25] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020. 3

[26] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran

Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[27] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020. 3

[28] Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22214–22224, 2023. 2, 6, 7, 8

[29] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *Advances in neural information processing systems*, 2021(DB1):1, 2021. 6, 8

[30] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022. 2

[31] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21 (146):1–76, 2020. 3

[32] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022. 2, 4, 6, 7, 8

[33] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 8

[34] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017. 3

[35] Ya Sun, Sijie Mai, and Haifeng Hu. Learning to balance the learning rates between various modalities via adaptive tracking factor. *IEEE Signal Processing Letters*, 28:1650–1654, 2021. 2

[36] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[37] LCM The, Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024. 2

[38] Valentin Thomas, Fabian Pedregosa, Bart Merriënboer, Pierre-Antoine Manzagol, Yoshua Bengio, and Nicolas Le Roux. On the interplay between noise and curvature and its effect on optimization and generalization. In *International Conference on Artificial Intelligence and Statistics*, pages 3503–3513. PMLR, 2020. 3

[39] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (11), 2008. 7, 8, 2

[40] Dong Wang, Di Hu, Xingjian Li, and Dejing Dou. Temporal relational modeling with self-supervision for action segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2729–2737, 2021. 1

[41] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12695–12705, 2020. 2, 6

[42] Yake Wei, Weiran Shen, and Di Hu. Mmpareto: Innocent uni-modal assistance for enhanced multi-modal learning. 2, 6

[43] Yake Wei, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. *arXiv preprint arXiv:2407.09705*, 2024. 2

[44] Torsten N Wiesel. Postnatal development of the visual cortex and the influence of environment. *Nature*, 299(5884):583–591, 1982. 1

[45] Torsten N Wiesel and David H Hubel. Single-cell responses in striate cortex of kittens deprived of vision in one eye. *Journal of neurophysiology*, 26(6):1003–1017, 1963. 3

[46] 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 *International Conference on Machine Learning*, pages 24043–24055. PMLR, 2022. 6

[47] Shaoxuan Xu, Menglu Cui, Chengxiang Huang, Hongfa Wang, et al. Balancebenchmark: A survey for imbalanced learning. *arXiv preprint arXiv:2502.10816*, 2025. 2

[48] Gang Yan, Hao Wang, and Jian Li. Critical learning periods in federated learning. *arXiv preprint arXiv:2109.05613*, 2021. 3

[49] Gang Yan, Hao Wang, Xu Yuan, and Jian Li. Criticalfl: A critical learning periods augmented client selection framework for efficient federated learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2898–2907, 2023. 3

[50] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arxiv 2016. *arXiv preprint arXiv:1606.06259*. 6

Adaptive Unimodal Regulation
for Balanced Multimodal Information Acquisition

Supplementary Material

A. Orthogonality proof

This proof supports the analysis presented in Section 3.3.2, specifically Equation 14, where the regulation term $P_{m;b}^t$ involves gradients $\mathbf{g}_k^t$ from multiple batches. The analysis assumes that, due to the high dimensionality of the space, the gradients $\mathbf{g}_k^t$ from different batches are nearly orthogonal. Here, we formally prove this assumption by showing that random vectors sampled from the surface of a high-dimensional hypersphere are nearly orthogonal with high probability.

Lemma 1. In high-dimensional spaces, let $\mathbf{g}_z^t, \mathbf{g}_k^t \in \mathbb{R}^n$ be two random vectors uniformly sampled from the surface of an n-dimensional hypersphere with magnitudes $\|\mathbf{g}_z^t\| = a$ and $\|\mathbf{g}_k^t\| = b$. As $n \to \infty$, these vectors are nearly orthogonal with high probability. Specifically, their dot product satisfies:

$$\mathbf{g}_z^t \cdot \mathbf{g}_k^t = ab \cos\theta \approx 0. \qquad (17)$$

Proof of Lemma 1. Let $\mathbf{g}_z^t$ and $\mathbf{g}_k^t$ be two random vectors in $\mathbb{R}^n$ with magnitudes $\|\mathbf{g}_z^t\| = a$ and $\|\mathbf{g}_k^t\| = b$. The dot product is given by:

$$\mathbf{g}_z^t \cdot \mathbf{g}_k^t = ab \cos\theta. \qquad (18)$$

To analyze the distribution of the angle θ in high-dimensional space, we consider the geometry of the n-dimensional unit hypersphere. Any vector $\mathbf{x} \in \mathbb{R}^n$ with unit norm, i.e., $\|\mathbf{x}\|_2 = 1$, lies on the surface of the unit hypersphere. It can be parameterized in spherical coordinates as:

$$\mathbf{x} = (x_1, x_2, \ldots, x_n), \quad \text{where } x_i \in \mathbb{R}, \ \sum_{i=1}^n x_i^2 = 1. \ (19)$$

The components of $\mathbf{x}$ in spherical coordinates are:

$$\begin{aligned}
x_1 &= \cos\phi_1, \\
x_2 &= \sin\phi_1 \cos\phi_2, \\
x_3 &= \sin\phi_1 \sin\phi_2 \cos\phi_3, \\
&\vdots \\
x_n &= \prod_{i=1}^{n-1} \sin\phi_i,
\end{aligned} \qquad (20)$$

where $\phi_1, \phi_2, \ldots, \phi_{n-2} \in [0, \pi]$, and $\phi_{n-1} \in [0, 2\pi]$. The surface element of the hypersphere is:

$$dS = (\sin\phi_1)^{n-2}(\sin\phi_2)^{n-3} \cdots \sin\phi_{n-2}\, d\phi_1 d\phi_2 \cdots d\phi_{n-1}. \qquad (21)$$

Without loss of generality, let one vector $\mathbf{g}_z^t$ be fixed along the x_1-axis, $\mathbf{g}_z^t = (a, 0, \ldots, 0)$. The second vector $\mathbf{g}_k^t$ can be parameterized using spherical coordinates. The angle θ between $\mathbf{g}_z^t$ and $\mathbf{g}_k^t$ is the same as ϕ_1, the first coordinate angle, so:

$$\cos\phi_1 = \cos\theta. \qquad (22)$$

The relevant term in the hypersphere surface element is:

$$p_n(\phi_1) \propto (\sin\phi_1)^{n-2}. \qquad (23)$$

This shows that the probability density of ϕ_1 (or θ) depends on the sine function raised to the power of $(n-2)$. For large n, $(\sin\phi_1)^{n-2}$ is sharply concentrated around $\phi_1 = \pi/2$ because $\sin\phi_1$ reaches its maximum at $\pi/2$. As $n \to \infty$, this concentration becomes stronger, leading to $\phi_1 \approx \frac{\pi}{2}$ with high probability. Since $\phi_1 \approx \pi/2$, we have:

$$\cos\phi_1 = \cos\theta \approx 0. \qquad (24)$$

Thus, in high-dimensional spaces, the angle θ between two random vectors concentrates around $\pi/2$, leading to:

$$\mathbf{g}_z^t \cdot \mathbf{g}_k^t = ab \cos\theta \approx 0. \qquad (25)$$

This demonstrates that the vectors are nearly orthogonal as $n \to \infty$.

B. Gradient norm analysis

This section aims to demonstrate that the regulation term $P_{m;b}^t$, introduced to regulate the information-sufficient modalities during the prime learning window, does not hinder the convergence of the optimization process. Specifically, we analyze the gradient norm and show that, under proper parameter settings, the convergence rate remains consistent with that of the original optimization objective without the regulation term.

Lemma 2. At training epoch t and batch b, consider the optimization objective:

$$\mathcal{L}(w_{m;b}^t) = \mathcal{L}_{joint}(w_{m;b}^t) + P_{m;b}^t, \qquad (26)$$

where $\mathcal{L}_{joint}(w_{m;b}^t)$ is the multimodal joint loss function, and the regulation term $P_{m;b}^t$ is defined as:

$$P_{m;b}^t = \frac{\alpha\eta^2}{2} \sum_{k=0}^{b} \|g_k^t\|^2. \qquad (27)$$

Here, $\alpha > 0$ is the regularization coefficient, $\eta > 0$ is the learning rate, and g_k^t denotes the gradient of batch k at epoch t. If α and η are sufficiently small, the convergence rate remains of the same order as without the regulation term.

Proof of Lemma 2. During the training, the weight update rule is given by:

$$w_{m;b+1}^t = w_{m;b}^t - \eta\nabla\mathcal{L}(w_{m;b}^t), \qquad (28)$$

where:

$$\nabla\mathcal{L}(w_{m;b}^t) = \nabla\mathcal{L}_{joint}(w_{m;b}^t) + \nabla P_{m;b}^t. \qquad (29)$$

The gradient of the regulation term $P_{m;b}^t$ is given by:

$$\nabla P_{m;b}^t = \alpha\eta^2 \sum_{k=0}^{b} g_k^t. \qquad (30)$$

Assuming that $\mathcal{L}(w)$ is L-Lipschitz smooth, we have:

$$\mathcal{L}(w_{m;b+1}^t) \leq \mathcal{L}(w_{m;b}^t) + \nabla\mathcal{L}(w_{m;b}^t)^T(w_{m;b+1}^t - w_{m;b}^t)$$
$$+ \frac{L}{2}\|w_{m;b+1}^t - w_{m;b}^t\|^2. \qquad (31)$$

Substituting $w_{m;b+1}^t - w_{m;b}^t = -\eta\nabla\mathcal{L}(w_{m;b}^t)$, we obtain:

$$\mathcal{L}(w_{m;b+1}^t) \leq \mathcal{L}(w_{m;b}^t) - \eta\|\nabla\mathcal{L}(w_{m;b}^t)\|^2 + \frac{L\eta^2}{2}\|\nabla\mathcal{L}(w_{m;b}^t)\|^2. \qquad (32)$$

The gradient norm is expressed as:

$$\|\nabla\mathcal{L}(w_{m;b}^t)\|^2 = \|\nabla\mathcal{L}_{joint}(w_{m;b}^t)\|^2$$
$$+ 2\nabla P_{m;b}^t \cdot \nabla\mathcal{L}_{joint}(w_{m;b}^t) \qquad (33)$$
$$+ \|\nabla P_{m;b}^t\|^2.$$

Due to the high dimensionality of the space, as demonstrated in Section A, the regulation term gradient $\nabla P_{m;b}^t$ and the joint loss gradient $\nabla\mathcal{L}_{joint}(w_{m;b}^t)$ are nearly orthogonal. As a result, their dot product can be approximated as:

$$\nabla P_{m;b}^t \cdot \nabla\mathcal{L}_{joint}(w_{m;b}^t) \approx 0. \qquad (34)$$

The gradient of the regulation term is bounded as:

$$\|\nabla P_{m;b}^t\| = \alpha\eta^2 \left\|\sum_{k=0}^{b} g_k^t\right\| \leq \alpha\eta^2 bG, \qquad (35)$$

where G is the upper bound of the gradient norm $\|g_k^t\|$. Thus, the term satisfies:

$$\|\nabla\mathcal{L}(w_{m;b}^t)\|^2 \leq \|\nabla\mathcal{L}_{joint}(w_{m;b}^t)\|^2 + \alpha^2\eta^4 b^2 G^2. \qquad (36)$$

For sufficiently small α and η, the additional term $\alpha^2\eta^4 b^2 G^2$ becomes negligible. Therefore, the convergence rate remains of the same order as without $P_{m;b}^t$.

C. Supplementary t-SNE analysis

(a) InfoReg*.

(b) Joint training.

(c) Joint training*.
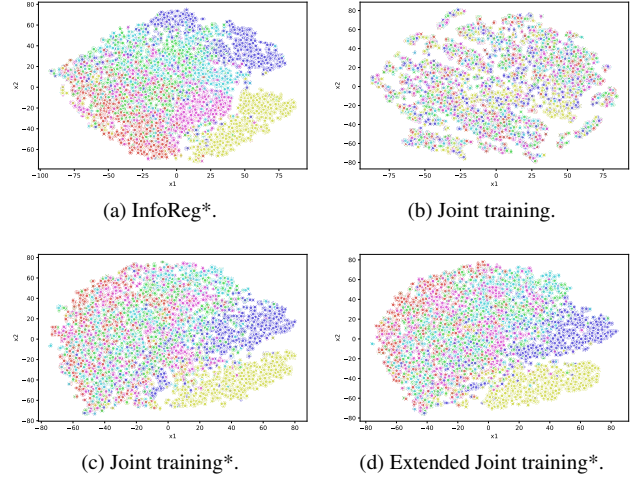
(d) Extended Joint training*.

Figure 9. The representations of the video modality on CREMA-D by t-SNE [39] across different methods are shown. InfoReg* and Joint training* denote InfoReg and Joint training with unimodal loss respectively. "Extended Joint training*" denotes Joint training* that is extended to 100 epochs.

To provide a more comprehensive evaluation of the proposed InfoReg method, we extend our analysis by incorporating t-SNE visualizations of video modality representations for InfoReg* and Joint training* on the CREMA-D dataset. Here, InfoReg* denotes InfoReg with unimodal loss, and Joint training* denotes Joint training with unimodal loss. As shown in Figure 9, InfoReg* and Joint training* learn better representations than Joint training. This is because the unimodal loss helps the multimodal model acquire more information. Additionally, the features learned by Joint training* and Extended Joint training* are similar, as shown in Figure 9c and Figure 9d. This indicates that extending the training time cannot compensate for the lack of information acquired during the prime learning window. Furthermore, InfoReg* learns better representations than both Joint training* and Extended Joint training*. This demonstrates that, with unimodal loss, our method can still help information-insufficient modalities acquire more information in the prime learning window. As a result, InfoReg* learns better representation.
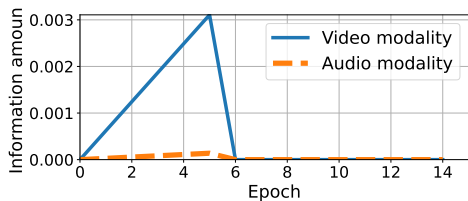
D. Supplementary experiments

Figure 10. Violence Flow dataset example, showcasing video modality dominance.

Dataset	Violence Flow	Hateful Memes
Joint training	89.21	55.00
InfoReg	**90.56**	**56.20**

Table 7. Accuracy comparison.

To further evaluate the effectiveness of InfoReg under diverse dataset conditions, we conducted experiments on the Violence Flow [14] and Hateful Memes [25] datasets. These datasets present different challenges: Violence Flow emphasizes anomaly detection, where the video modality quickly becomes dominant, while Hateful Memes requires cooperation between modalities due to its complex multi-modal nature.

Figure 10 illustrates the information amount during training on the Violence Flow, where the video modality demonstrates dominance during the prime learning window. InfoReg can identify this dominant modality.

The Hateful Memes dataset requires significant cooperation between modalities. As shown in the Table 7, Despite the increased complexity, InfoReg can still improve the performance of the model.