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 $P^t_{m;b}$ involves gradients $\mathbf{g}^t_k$ from multiple batches. The analysis assumes that, due to the high dimensionality of the space, the gradients $\mathbf{g}^t_k$ 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 $\mathbf{g}^t_z, \mathbf{g}^t_k \in \mathbb{R}^n$ be two random vectors uniformly sampled from the surface of an n-dimensional hypersphere with magnitudes $\|\mathbf{g}^t_z\| = a$ and $\|\mathbf{g}^t_k\| = b$. As $n \to \infty$, these vectors are nearly orthogonal with high probability. Specifically, their dot product satisfies:

$$\mathbf{g}^t_z \cdot \mathbf{g}^t_k = ab\cos\theta \approx 0. \tag{17}$$

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

$$\mathbf{g}^t_z \cdot \mathbf{g}^t_k = ab\cos\theta. \tag{18}$$

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:

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

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

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

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:

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

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

$$\cos\phi_1 = \cos\theta. \tag{22}$$

The relevant term in the hypersphere surface element is:

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

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:

$$\cos\phi_1 = \cos\theta \ \approx 0. \tag{24}$$

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

$$\mathbf{g}^t_z \cdot \mathbf{g}^t_k = ab\cos\theta \approx 0. \tag{25}$$

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 $P^t_{m;b}$, 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:

$$\mathcal{L}(w^t_{m;b}) = \mathcal{L}_{joint}(w^t_{m;b}) + P^t_{m;b}, \tag{26}$$

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:

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

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:

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

where:

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

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

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

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

$$\begin{aligned}\mathcal{L}(w_{m;b+1}^t) \le \mathcal{L}(w_{m;b}^t) + \nabla\mathcal{L}(w_{m;b}^t)^T(w_{m;b+1}^t - w_{m;b}^t) \\ + \frac{L}{2}\|w_{m;b+1}^t - w_{m;b}^t\|^2.\end{aligned} \tag{31}$$

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

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

The gradient norm is expressed as:

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

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:

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

The gradient of the regulation term is bounded as:

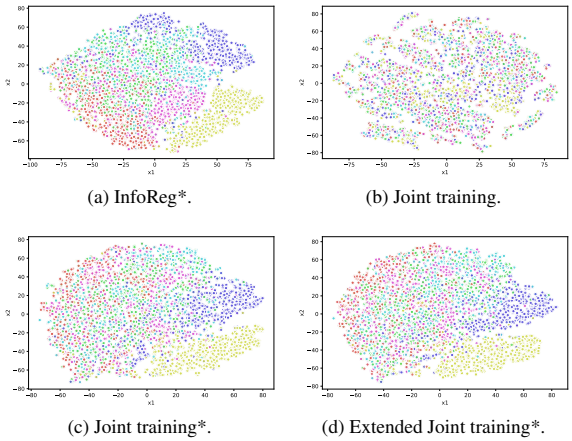
$$\|\nabla P_{m;b}^t\| = \alpha\eta^2\left\|\sum_{k=0}^{b} g_k^t\right\| \le \alpha\eta^2 bG, \tag{35}$$

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

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

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

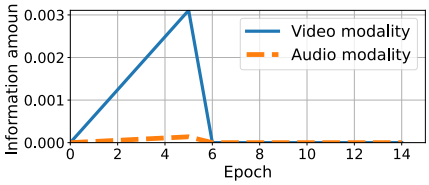


Figure 10. Violence Flow dataset example, showcasing video modality dominance.

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 multimodal 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 perforemance of the model.