

# Appendix

## A PROOF OF THEOREM 1

**Theorem 1.** Suppose there exists encoders  $h_p : \mathbf{X}^{(p)} \rightarrow \mathbf{Z}^{(p)}$  and  $h_q : \mathbf{X}^{(q)} \rightarrow \mathbf{Z}^{(q)}$ , such that  $\mathbf{Z}^{(p)} = \mathbf{Z}^{(q)}$ . And the fusion functions  $g_x$  and  $g_z$ , applying to  $\{\mathbf{X}^{(v)}\}_{v=p,q}$  and  $\{\mathbf{Z}^{(v)}\}_{v=p,q}$  respectively, allow the fused features to retain maximum information content. Then encoders  $h_p$  and  $h_q$  would disregard the view gap, resulting in information loss:

$$I(g_x(\mathbf{X}^{(p)}, \mathbf{X}^{(q)}); \mathbf{Y}^*) - I(g_z(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}); \mathbf{Y}^*) \geq \delta_{pq}. \quad (1)$$

**Proof:** Referring to [2], we provide a similar proof. Contemplate the joint mutual information  $I(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}; \mathbf{Y}^*)$ . Utilizing the chain rule, we can decompose it as follows:

$$\begin{aligned} I(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}; \mathbf{Y}^*) &= I(\mathbf{Z}^{(p)}; \mathbf{Y}^*) + I(\mathbf{Z}^{(q)}; \mathbf{Y}^* | \mathbf{Z}^{(p)}) \\ &= I(\mathbf{Z}^{(q)}; \mathbf{Y}^*) + I(\mathbf{Z}^{(p)}; \mathbf{Y}^* | \mathbf{Z}^{(q)}). \end{aligned} \quad (2)$$

Suppose  $\mathbf{Z}^{(p)}$  and  $\mathbf{Z}^{(q)}$  are perfect aligned, namely  $\mathbf{Z}^{(p)} = \mathbf{Z}^{(q)}$ , then

$$I(\mathbf{Z}^{(q)}; \mathbf{Y}^* | \mathbf{Z}^{(p)}) = I(\mathbf{Z}^{(p)}; \mathbf{Y}^* | \mathbf{Z}^{(q)}) = 0, \quad (3)$$

therefore, it can be concluded that

$$I(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}; \mathbf{Y}^*) = I(\mathbf{Z}^{(p)}; \mathbf{Y}^*) = I(\mathbf{Z}^{(q)}; \mathbf{Y}^*). \quad (4)$$

According to the information bottleneck theorem [1],

$$I(\mathbf{Z}^{(p)}; \mathbf{Y}^*) \leq I(\mathbf{X}^{(p)}; \mathbf{Y}^*), \quad I(\mathbf{Z}^{(q)}; \mathbf{Y}^*) \leq I(\mathbf{X}^{(q)}; \mathbf{Y}^*). \quad (5)$$

Therefore, the subsequent sequence of inequalities is valid:

$$\begin{aligned} I(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}; \mathbf{Y}^*) &= \min\{I(\mathbf{Z}^{(p)}; \mathbf{Y}^*), I(\mathbf{Z}^{(q)}; \mathbf{Y}^*)\} \\ &\leq \min\{I(\mathbf{X}^{(p)}; \mathbf{Y}^*), I(\mathbf{X}^{(q)}; \mathbf{Y}^*)\} \\ &\leq \max\{I(\mathbf{X}^{(p)}; \mathbf{Y}^*), I(\mathbf{X}^{(q)}; \mathbf{Y}^*)\} \\ &\leq I(\mathbf{X}^{(p)}, \mathbf{X}^{(q)}; \mathbf{Y}^*), \end{aligned} \quad (6)$$

where the final inequality is derived from the observation that when view  $p$  and  $q$  are trained to be completely consistent, information loss occurs. Therefore, due to the perfect fusion functions  $g_x$  and  $g_z$ , allowing the fused features to retain maximum information content, we have

$$\begin{aligned} I(g_x(\mathbf{X}^{(p)}, \mathbf{X}^{(q)}); \mathbf{Y}^*) - I(g_z(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}); \mathbf{Y}^*) \\ &= I(\mathbf{X}^{(p)}, \mathbf{X}^{(q)}; \mathbf{Y}^*) - I(\mathbf{Z}^{(p)}, \mathbf{Z}^{(q)}; \mathbf{Y}^*) \\ &\geq \max\{I(\mathbf{X}^{(p)}; \mathbf{Y}^*), I(\mathbf{X}^{(q)}; \mathbf{Y}^*)\} - \min\{I(\mathbf{X}^{(p)}; \mathbf{Y}^*), I(\mathbf{X}^{(q)}; \mathbf{Y}^*)\} \\ &= \delta_{pq}. \end{aligned} \quad (7)$$

## B PROOF OF THEOREM 2

**Theorem 2.** Imposing the constraints of  $\ell_{sp}(p, q)$  can prevent the convergence of representations from two views into uniformity, i.e.  $\hat{\mathbf{Z}}^{(p)} \neq \hat{\mathbf{Z}}^{(q)}$ , thereby preserving the diverse information present across the multiple views.

**Proof:** The loss  $\ell_{sp}(p, q)$  is defined as

$$\ell_i^{(p)} = -\log \frac{\exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})/\tau_l)}{\sum_{j=1}^N \left[ \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(p)})/\tau_l) + \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(q)})/\tau_l) \right]}, \quad (8)$$

$$\ell_{sp}(p, q) = \sqrt{\frac{1}{2N} \sum_{v=p,q} \sum_{i=1}^N (\hat{\ell}_i^{(v)} - \hat{\ell}_\mu)^2}, \quad (9)$$

where  $\hat{\ell}_\mu = \frac{1}{2N} \sum_{i=1}^N (\hat{\ell}_i^{(p)} + \hat{\ell}_i^{(q)})$ . Hence, the purpose of  $\ell_{sp}(p, q)$  is to minimize the variance of the contrastive losses incurred by any sample in the  $p$ -th view and the  $q$ -th view. We use the symbol " $\rightarrow$ " to signify convergence. Therefore, the  $\ell_{sp}(p, q) \rightarrow 0$  is equivalent to  $\hat{\ell}_i^{(v)} \rightarrow \hat{\ell}_\mu$ , ( $v = p, q$ ).

We make an approximation to  $\hat{\ell}$ , with

$$\hat{\ell}_\mu \approx \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i^{(p)} \approx \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i^{(q)}. \quad (10)$$

Then the convergence of loss function  $\ell_{sp}(p, q) \rightarrow 0$  can be approximated to

$$\begin{cases} \hat{\ell}_i^{(p)} \rightarrow \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i^{(p)}, \\ \hat{\ell}_i^{(q)} \rightarrow \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i^{(q)}. \end{cases} \quad (11)$$

When convergence  $\hat{\ell}_i^{(p)} \rightarrow \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i^{(p)}$  reaches its limit, there exists

$$\begin{cases} \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})/\tau_l) \rightarrow \frac{1}{N} \sum_{i=1}^N \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})/\tau_l), \\ \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(p)})/\tau_l) \rightarrow \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(p)})/\tau_l), \\ \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(q)})/\tau_l) \rightarrow \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(q)})/\tau_l). \end{cases} \quad (12)$$

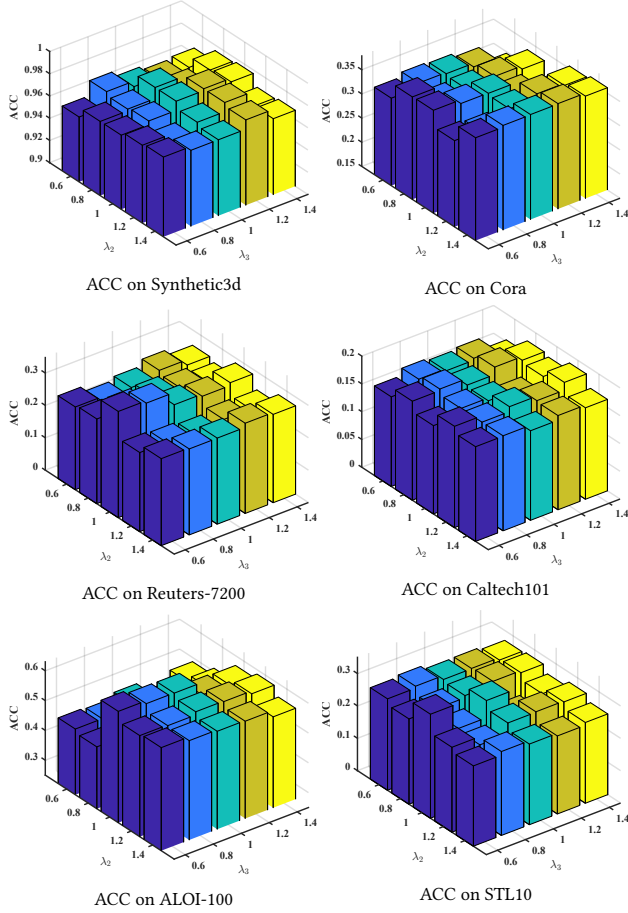
In the above expression, as the loss converges, the right side of the arrow approaches a constant. For the sake of convenience in representation, we denote

$$\begin{cases} \frac{1}{N} \sum_{i=1}^N \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})/\tau_l) = A, \\ \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(p)})/\tau_l) = B_1, \\ \frac{1}{N} \sum_{i=1}^N \sum_{j=1}^N \exp(f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_j^{(q)})/\tau_l) = B_2. \end{cases} \quad (13)$$

Since  $\hat{\ell}_i^{(p)} \rightarrow \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i^{(p)}$ , then we have

$$\min\{\hat{\ell}_i^{(p)}\}_{i=1}^N \leq \hat{\ell}_i^{(p)} \leq \max\{\hat{\ell}_i^{(p)}\}_{i=1}^N, \quad (14)$$

where if we regard the denominator of  $\hat{\ell}_i^{(p)}$  as a constant value ( $B_1 + B_2$ ), then we can obtain the final approximate inequality:



**Figure 1: Sensitivity analysis of the hyper-parameters on six datasets.**

$$\min\{f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})\}_{i=1}^N \leq f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)}) \leq \max\{f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})\}_{i=1}^N. \quad (15)$$

Therefore, it is evident that there exists a sample  $\hat{\mathbf{z}}_i$  such that  $f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)}) < \max\{f(\hat{\mathbf{z}}_i^{(p)}, \hat{\mathbf{z}}_i^{(q)})\}_{i=1}^N \leq 1$ . Thus, we can draw the conclusion that  $\hat{\mathbf{Z}}^{(p)} \neq \hat{\mathbf{Z}}^{(q)}$  with the convergence of loss  $\ell_{sp}(p, q)$ .

## C FURTHER ABLATION STUDY

We demonstrate further ablation experiments on Cora, Caltech101 and STL10 datasets. The experimental results corroborate our analysis in the main text.

**Table 1: Ablation study on Cora, Caltech101 and STL10 datasets. M1, M2 and M3 are abbreviations for ShaSpec Module(M1), ShaTree Module(M2) and SpecTree Module(M3), respectively. ✓ denotes TGM-MVC with the Module.**

Datasets	M1	M2	M3	ACC	NMI	PUR
Cora	✓	✓	✓	<b>37.59</b>	<b>13.93</b>	<b>38.44</b>
	✓			32.13	8.250	33.53
	✓		✓	35.01	12.82	35.56
	✓	✓		35.86	10.52	36.15
Caltech101	✓	✓	✓	<b>20.31</b>	<b>40.62</b>	<b>37.33</b>
	✓			15.87	38.56	35.62
	✓		✓	15.39	34.16	32.46
	✓	✓		17.01	39.05	36.02
STL10	✓	✓	✓	<b>32.13</b>	<b>25.93</b>	<b>34.48</b>
	✓			28.08	26.07	30.43
	✓		✓	29.72	23.92	30.05
	✓	✓		27.87	25.60	28.38

## D FURTHER SENTIMENT ANALYSIS

In order to assess the robustness of our model, we conducted further sensitivity analysis on Synthetic3d, Cora, Reuters-7200, Caltech101 and STL10 datasets.

## REFERENCES

- [1] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep Variational Information Bottleneck. In *International Conference on Learning Representations*.
- [2] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. 2023. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7661–7671.