Open-set Cross Modal Generalization via Multimodal Unified Representation

Supplementary Material

The citation numbers are consistent with those in the main text.

1. Mask of FCMI

We also conducted an analysis on different masking strategies. As shown in Figure 1, applying the same mask to paired multimodal samples helps improve model performance. This approach facilitates more precise and detailed alignment between modalities, ensuring semantic consistency in the unmasked regions while applying the mask to the same positions across modalities. In contrast, using different masking positions for each modality in paired samples leads to a decline in performance, as it disrupts the semantic alignment across the modalities.

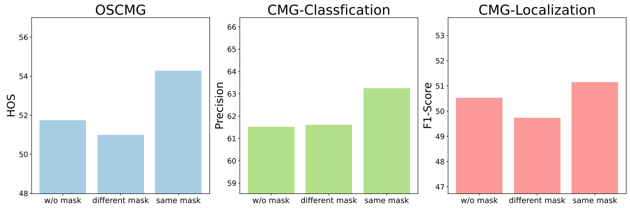


Figure 1. Experimental results of different Mask.

2. Codebook Size

The size of the representation space also affects the model's performance. As shown in Figure 2, we experimented with five different settings: 256, 400, 512, 800, and 1024. Among these, 400 led by a significant margin over the other settings. Therefore, we chose a codebook size of 400 as the final setting for our model.

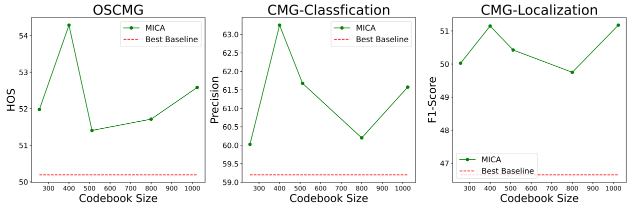


Figure 2. Experimental results of different Codebook Size.

3. Ablation on CMG

The experimental results of Table 1 and Table **??** are similar. L_{coarse} serves as the foundation of the model, while L_{fine} and L_{cujp} further refine the unified representation space and enhance the model's open-domain detection capabilities.

L_{fine}	L_{coarse}	L_{cujp}	AVE		AVVP		AVE→AVVP		UCF(v)↔VGG(a)	
			V→A	A→V	V→A	A→V	V→A	A→V	V→A	A→V
✓	-	-	7.1	5.2	13.4	13.7	15.9	7.4	10.5	8.2
-	✓	-	54.3	55.2	39.6	37.8	50.5	46.3	70.3	61.7
-	-	✓	5.6	5.1	0	6.0	0	0	13.0	9.7
✓	✓	-	**56.1**	57.0	38.9	35.8	52.2	43.3	70.8	**64.6**
✓	-	✓	6.4	4.8	13.4	13.7	15.9	7.4	11.1	8.2
-	✓	✓	53.8	52.4	43.8	45.9	**56.7**	**54.9**	67.4	62.3
✓	✓	✓	**56.1**	**57.1**	**45.2**	**48.2**	56.3	**54.9**	**75.3**	64.5

Table 1. Ablation study of the three losses proposed by our model on CMG.

4. Computational Efficiency

As shown in Table 2, compared to CMCM [3] and DCID [4], our method requires more GPU memory and longer per-epoch training time, but achieves better performance, reflecting a trade-off between performance and resources. CUJP8, despite having more split block reordering, optimizes memory usage and reduces training time compared to MMJP6 [2]. Increasing the number of splits (CUJP4 vs. CUJP8) leads to higher memory usage but better performance in multimodal alignment. CMCM requires more epochs due to warm-start techniques. Inference time differences across all models are minimal and task-dependent. For reproducibility, the complete source code is provided in the supplementary materials.

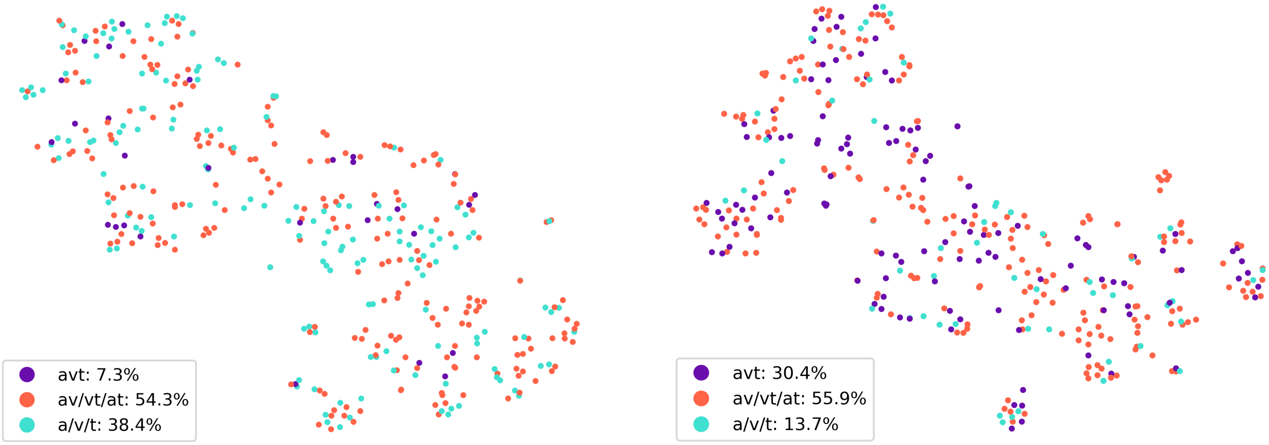
Method	GPU Memory Usage	Time per Epoch	Total Epochs	OSCMG Avg.	CMG Avg.
CMCM	6.25GB	1.41h	8	44.47	44.78
DCID	7.90GB	1.72h	5	50.19	52.93
MICU (MMJP6)	14.77GB	2.30h	5	52.00	52.46
MICU (CUJP4)	9.07GB	2.13h	5	52.56	53.75
MICU (CUJP8)	13.30GB	2.22h	5	54.29	57.20

Table 2. Comparison of computational efficiency with the original backbone (batch size: 80, GPU: RTX 3090).

5. Unified Representation Space Visualization

As shown in Figure 3, the two subfigures illustrate the representation spaces of DCID [4] after pre-training and our proposed model. The visualization maps audio-video-text triplets from the Valor32K dataset [1] into the unified representation space (codebook). Codewords quantized by all three modalities with a proportion of $\geq$10% are marked in purple, those shared by any two modalities with $\geq$10% appear in orange, while those dominated by a single modality are shown in cyan. The bottom left of the figure indicates the proportion of each color.

A higher proportion of cyan suggests an imbalanced multimodal distribution, indicating larger modality discrepancies, whereas more purple signifies stronger cross-modal alignment, aligning with the goal of a unified representation. As observed, our model achieves significantly better multimodal integration compared to DCID.



(a) DCID Representation Space Visualization

(b) MICU Representation Space Visualization

Figure 3. Purple (avt) indicates where all three modalities have quantized activations $\geq$10%, orange (av/vt/at) for two modalities, and cyan (a/v/t) for a single modality.

References

[1] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. *arXiv preprint arXiv:2304.08345*, 2023. 1

[2] Hao Dong, Eleni Chatzi, and Olga Fink. Towards multimodal open-set domain generalization and adaptation through self-supervision. *arXiv preprint arXiv:2407.01518*, 2024. 1

[3] Alexander H Liu, SouYoung Jin, Cheng-I Jeff Lai, Andrew Rouditchenko, Aude Oliva, and James Glass. Cross-modal discrete representation learning. *arXiv preprint arXiv:2106.05438*, 2021. 1

[4] Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36, 2024. 1