Bridging Domain Generalization to Multimodal Domain Generalization via Unified Representations

Supplementary Material

1. Inplementation Details

To ensure fair experimental comparisons, we adopt the same modality backbones as SimMMDG [5] and CMRF [6], with our experimental setup based on the MMAction2 toolkit [4]. The feature dimensions for video, audio, and optical flow are 2304, 512, and 2048, respectively. In constructing the unified representation, both the general information and specific information maintain a consistent dimension of 512 across all modalities. Each modality's general encoder and specific encoder consist of a two-layer MLP with an input dimension matching the respective modality's feature dimension (2304, 512, or 2048), a hidden layer of size 2048, and an output dimension of 512. The scalar temperature parameter τ is set to 0.1.

For modality-specific encoders, we use the SlowFast network's slow-only pathway for optical flow encoding, initialized with Kinetics-400 pre-trained weights, the SlowFast network [7] for visual encoding, also initialized with Kinetics-400 pre-trained weights [9], and ResNet-18 [8] for audio encoding, initialized with weights from the VGGSound pre-trained checkpoint [3]. The hyperparameters $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are set to 1.0, 2.0, 2.0, and 1.0, respectively. For UR-Mixup, the Beta distribution parameter α is set to 0.2, while for UR-JiGen, the Jigsaw number P is set to 256.

All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU, with training performed for 20 epochs. UR-IBN and UR-Mixup require approximately 3 hours, while UR-JiGen takes around 3.5 hours. In our implementation, we initially experimented with a warm-start approach, where training was conducted in two phases: first, learning the unified representation, followed by applying the DG method. However, experimental results indicated that warm-start had no significant impact. Consequently, we adopted a more efficient end-to-end training strategy, where the construction of the unified representation and the application of the DG method occur simultaneously.

2. More Experiments

More Experiments about Multi-modal single-source DG: As shown in Table 1, directly transferring DG methods [2, 11, 13] to MMDG results in significantly inferior performance compared to models specifically designed for MMDG [5, 6]. In contrast, our proposed approach substantially improves their performance in the MMDG setting. Notably, UR-JiGen and UR-Mixup demonstrate competitive results against previous state-of-the-art models, further validating the effectiveness of our method.

Method	Modality			EPIC-Kitchens	HAC
	Video	Audio	Flow		
Base	✓	✓		52.34	54.18
RNA-Net [12]	✓	✓		51.25	54.51
SimMMDG [5]	✓	✓		54.84	58.75
CMRF [6]	✓	✓		**57.64**	60.87
IBN [11]	✓	✓		51.43	53.62
JiGen [2]	✓	✓		52.69	55.15
Mixup [13]	✓	✓		53.28	56.29
UR-IBN (ours)	✓	✓		53.89	57.43
UR-JiGen (ours)	✓	✓		56.94	60.48
UR-Mixup (ours)	✓	✓		57.39	**61.32**
Base	✓		✓	53.64	56.80
RNA-Net [12]	✓		✓	53.86	57.26
SimMMDG [5]	✓		✓	57.32	60.63
CMRF [6]	✓		✓	59.26	62.45
IBN [11]	✓		✓	54.06	56.62
JiGen [2]	✓		✓	55.64	58.23
Mixup [13]	✓		✓	55.97	59.24
UR-IBN (ours)	✓		✓	55.97	58.35
UR-JiGen (ours)	✓		✓	60.21	62.44
UR-Mixup (ours)	✓		✓	**60.46**	**62.93**
Base		✓	✓	48.68	44.26
RNA-Net [12]		✓	✓	49.69	42.72
SimMMDG [5]		✓	✓	53.27	47.28
CMRF [6]		✓	✓	56.46	49.96
IBN [11]		✓	✓	49.35	44.86
JiGen [2]		✓	✓	51.87	46.48
Mixup [13]		✓	✓	52.33	46.72
UR-IBN (ours)		✓	✓	52.23	47.51
UR-JiGen (ours)		✓	✓	**56.70**	50.74
UR-Mixup (ours)		✓	✓	56.22	**51.29**

Table 1. Multi-modal **single-source** DG with different modalities on EPIC-Kitchens and HAC dataset.

More Experiments about Uni-modal performance in MMDG: As shown in Table 2, the results are consistent with those in Table 2, further reinforcing our previous findings.

More Experiments about recent methods: Here, we provide additional results of recent methods with unified representations. Specifically, we report the performance of mDSDI [1] and RDM [10] under the multimodal multi-source domain generalization setting.

For space efficiency, we report average results across domain splits on EPIC-Kitchens and HAC. As shown in Table 3, integrating our unified representation (UR) design with these methods yields consistent and substantial improvements, outperforming the prior SOTA (CMRF) across all modality combinations.

Method	Video	Audio	Video-Audio	Video	Flow	Video-Flow	Audio	Flow	Audio-Flow
Base (M1)	58.73	-	-	58.73	-	-	40.04	-	-
Base (M2)	-	40.04	-	-	58.30	-	-	58.30	-
Base (MM)	56.65	38.62	59.63	55.28	55.78	60.89	39.42	54.86	53.14
JiGen (M1)	61.60	-	-	61.60	-	-	42.72	-	-
JiGen (M2)	-	42.72	-	-	60.77	-	-	60.77	-
JiGen (MM)	58.98	40.67	61.08	57.14	56.64	61.79	40.26	56.38	58.93
Mixup (M1)	61.92	-	-	61.92	-	-	43.74	-	-
Mixup (M2)	-	43.74	-	-	60.89	-	-	**60.89**	-
Mixup (MM)	58.52	39.31	61.18	57.86	57.24	62.08	40.38	57.08	58.32
UR-JiGen (ours)	62.02	43.41	63.63	61.26	61.08	64.15	43.52	**60.89**	**63.32**
UR-Mixup (ours)	**62.45**	**43.79**	**64.77**	**62.34**	**61.51**	**66.42**	**44.24**	60.24	62.63

Table 2. The average results of uni-modal performance comparison under multi-modal multi-source DG on EPIC-Kitchens with 3 different modality combinations. M1, M2, and MM denote training settings where the data correspond to the first and second single-modal cases, and the multi-modal case, respectively, following the column header order.

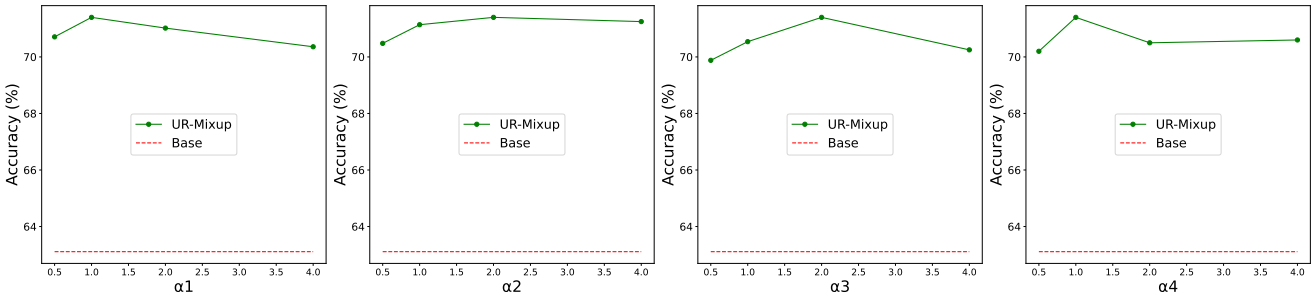

Figure 1. Parameter sensitivity analysis on HAC with video and audio data

Method	**Video+Audio**		**Video+Flow**		**Audio+Flow**		**Video+Audio+Flow**	
	EPIC	HAC	EPIC	HAC	EPIC	HAC	EPIC	HAC
CMRF	63.91	71.91	64.89	72.64	66.12	58.49	70.12	72.44
mDSDI	61.73	66.96	62.31	69.31	58.61	55.95	64.19	68.40
UR-mDSDI	65.61	73.65	67.25	73.76	65.99	**60.39**	71.25	74.42
RDM	62.04	67.58	62.64	69.87	58.67	56.37	63.93	68.61
UR-RDM	**66.18**	**74.17**	**68.07**	**74.29**	**66.48**	59.62	**71.88**	**74.79**

Table 3. Results of mDSDI and RDM with/without UR on EPIC-Kitchens and HAC.

3. Parameter Sensitivity Analysis

As shown in Figure 1, we conduct a comprehensive analysis of four loss hyperparameters in UR-Mixup by varying one parameter at a time while keeping the others fixed. Notably, our method exhibits minimal fluctuations across all parameter settings, indicating a lower sensitivity to hyperparameter selection.

4. Loss Function Ablation Study

As shown in Table 5 of the main paper, we ablate L_{scl} and L_{club}, where rows 1, 3, and 7 correspond to using L_{cls} only, $L_{\text{cls}}+L_{\text{scl}}$, and the full objective, respectively. We do not ablate L_{cls} since it is essential for classification. L_{rec} is meaningful only when L_{club} is used, as it ensures semantic completeness after disentanglement.

We further include results using $L_{\text{cls}}+L_{\text{club}}$, $L_{\text{cls}}+L_{\text{scl}}+L_{\text{club}}$, and $L_{\text{cls}}+L_{\text{club}}+L_{\text{rec}}$, in addition to the original settings. Each component contributes positively, confirming its utility in improving performance, as detailed in Table 4.

L_{cls}	L_{scl}	L_{club}	L_{rec}	D2,D3 → D1	D1,D3 → D2	D1,D2 → D3	Mean
✓				54.94	62.26	61.70	59.63
✓	✓			56.24	65.38	65.07	62.23
✓		✓		54.75	62.53	62.08	59.79
✓	✓	✓		56.44	67.41	67.25	63.70
✓		✓	✓	55.31	63.26	63.65	60.74
✓	✓	✓	✓	56.99	68.85	68.46	64.77

Table 4. Extended ablation study on loss components.

5. More Visualization

As shown in Figure 2, we provide additional visualizations of the learned embeddings. It can be observed that the general and specific information of each modality are well-separated and consistently aligned across domains. Fur-

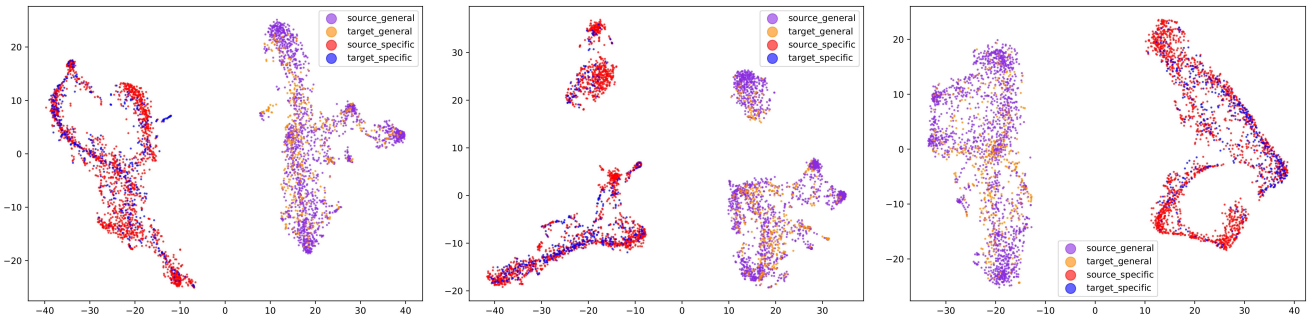

Figure 2. Visualization of the learned embeddings using t-SNE (D2, D3 D1 in EPIC-Kitchens for multi-modal multi-source DG).

thermore, the embeddings of flow, video, and audio exhibit strong alignment between the source and target domains.

References

[1] Manh-Ha Bui, Toan Tran, Anh Tran, and Dinh Phung. Exploiting domain-specific features to enhance domain generalization. *Advances in Neural Information Processing Systems*, 34:21189–21201, 2021. 1

[2] Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2229–2238, 2019. 1

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 1

[4] MMAction Contributors. Openmmlabs next generation video understanding toolbox and benchmark. 2020. 1

[5] Hao Dong, Ismail Nejjar, Han Sun, Eleni Chatzi, and Olga Fink. Simmmdg: A simple and effective framework for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 36:78674–78695, 2023. 1

[6] Yunfeng Fan, Wenchao Xu, Haozhao Wang, and Song Guo. Cross-modal representation flattening for multi-modal domain generalization. *Advances in Neural Information Processing Systems*, 37:66773–66795, 2025. 1

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 1

[8] 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. 1

[9] 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. 1

[10] Toan Nguyen, Kien Do, Bao Duong, and Thin Nguyen. Domain generalisation via risk distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2790–2799, 2024. 1

[11] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479, 2018. 1

[12] Mirco Planamente, Chiara Plizzari, Emanuele Alberti, and Barbara Caputo. Domain generalization through audio-visual relative norm alignment in first person action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1807–1818, 2022. 1

[13] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1