Decoupling Stability and Plasticity for Multi-Modal Test-Time Adaptation

Supplementary Material

This appendix contains supplementary results and detailed analyses that further validate our approach. The material is organized as follows:

- Sec. 1 offers detailed information regarding the benchmark and implementation.
- Sec. 2 provides additional evidence supporting our proposed redundancy score.
- Sec. 3 presents extended comparative experiments with recent MM-TTA methods.

1. More Experimental Details

1.1. Benchmarks

We construct two benchmarks based on Kinetics [4] and VGGSound [1], to evaluate the performance of state-of-the-art methods under multi-modal domain shifts during test-time adaptation. We introduce three experimental setups: uni-modal episodic corruption, uni-modal continual corruption, and interleaved modality continual corruption. As illustrated in Fig. 1, in the interleaved modality setup, corruption alternates continuously between different modalities (*e.g.* from video to audio).

Kinetics. The Kinetics dataset is a large, high-quality benchmark used to recognize human actions in videos. It includes about 500,000 video clips covering 600 different action classes, with each class having at least 600 clips. Each clip is approximately 10 seconds long and labeled with a single action. The videos were collected from YouTube. Our study focuses on a subset of the Kinetics dataset, which contains 50 action classes and 2,466 test pairs.

VGGSound. The VGGSound dataset is a large-scale benchmark for audio-visual correspondence. It contains short audio clips extracted from YouTube videos recorded "in the wild". This ensures a clear match between the audio and visual content, with the sound source being visually identifiable. Each video in the dataset is 10 seconds long. We have collected 14,046 visual-audio pairs for testing.

Kinetics50-C and VGGSound-C. Following privious work [8], we introduce 15 types of corruptions to the video modality data, including "Gaussian Noise", "Shot Noise", "Impulse Noise", "Defocus Blur", "Glass Blur", "Motion Blur", "Zoom Blur", "Snow", "Frost", "Fog", "Brightness", "Contrast", "Elastic Transform", "Pixelate", and "JPEG Compression". For the audio modality data, we introduce 6 types of corruptions, comprising "Gaussian Noise", "Paris Traffic Noise", "Crowd Noise", "Rainy Noise", "Thunder Noise" and "Windy Noise". Each corruption type is applied at five levels of severity.

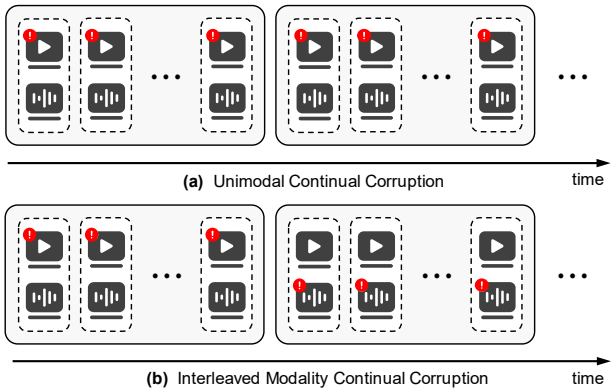

Figure 1. The illustration of uni-modal continual corruption and interleaved modality continual corruption.

1.2. Implementations

All baseline methods, along with our proposed approach, are optimized using the Adam optimizer, employing a learning rate of 1×10^{-4} for both Kinetics-C and VGGSound-C.

Tent. For Tent [7], following its official implementation[1], we set the tunable parameters to those within the LayerNorm modules.

EATA. For EATA [6], following its official implementation[2], we set the tunable parameters to those within the LayerNorm modules. Moreover, the exponential moving average (EMA) factor, cosine similarity threshold, and entropy threshold are set to 0.9, 0.05, and $0.4\ln(C)$, respectively, where C denotes the number of classes.

READ. For READ [8], following its official implementation[3], the tunable parameters are specified as the query, key, and value transformation matrices of the attention layer within the fusion block.

TSA. For TSA [2], following its official implementation[4], the modality-specific adapters are configured as the only tunable parameters.

DASP (Ours). Our proposed DASP adapts by updating two specialized modules: the stable and plastic adapters. The stable adapter uses a low-rank bottleneck design (rank $r = 32$), defined as $h = x + W_{up}\sigma(W_{down}x)$, which acts as a structural regularizer to prevent overfitting. In contrast, the plastic adapter is a full-rank linear module (rank

[1] https://github.com/DequanWang/tent
[2] https://github.com/mr-eggplant/EATA
[3] https://github.com/XLearning-SCU/2024-ICLR-READ
[4] https://github.com/chenmc1996/Uni-Modal-Distribution-Shift

$r = 768$), defined as $h = x + Wx$, providing enough capacity for feature realignment. Furthermore, we apply variance filtering by masking dimensions with near-zero variance ($\sigma^2 \approx 0$) before computing the redundancy score $R(\cdot)$, ensuring that invalid dimensions do not dilute the final score.

2. Further Analysis of the Redundancy Score

Theoretical Analysis. The distribution shift is modeled as a low-rank perturbation in the latent space. For a perturbed sample $\tilde{\mathbf{z}} \in \mathbb{R}^D$, we consider the dominant rank-1 component: $\tilde{\mathbf{z}} = \mathbf{z} + \alpha\mathbf{v}$. To formalize our analysis, we establish the following **Assumptions**:

1. The dimensions of $\mathbf{z}$ are centered and uncorrelated, *i.e.*, $\mathbb{E}[\mathbf{z}] = 0$ and $\mathrm{Cov}(\mathbf{z}) = \mathbf{I}_D$.

2. The shift intensity α is a random variable independent of $\mathbf{z}$, satisfying $\mathbb{E}[\alpha] = 0$ and $\mathrm{Var}(\alpha) = \sigma_\alpha^2 > 0$.

3. The shift direction $\mathbf{v}$ is non-sparse, such that $\|\mathbf{v}\|_0 \geq 2$.

Theorem. Let interdimensional redundancy be defined as $R(\mathbf{Z}) = \kappa \sum_{i \neq j} \mathcal{C}_{ij}^2$, where $\kappa = \frac{1}{D(D-1)} > 0$. Under Assumptions 1-3, the redundancy strictly increases under distribution shift, *i.e.*, $R(\tilde{\mathbf{Z}}) > R(\mathbf{Z}) = 0$.

Proof. By the independence in Assumption 2, the covariance matrix of perturbed features $\tilde{\mathbf{Z}}$ is given by:

$$\tilde{\Sigma} = \mathbb{E}[\tilde{\mathbf{z}}\tilde{\mathbf{z}}^\top] - \mathbb{E}[\tilde{\mathbf{z}}]\mathbb{E}[\tilde{\mathbf{z}}]^\top = \Sigma + \sigma_\alpha^2 \mathbf{v}\mathbf{v}^\top. \tag{1}$$

From Assumption 1, we have $\Sigma = \mathbf{I}_D$. Thus, for any $i \neq j$, the (i, j)-th entry of $\tilde{\Sigma}$ is:

$$\tilde{\Sigma}_{ij} = \Sigma_{ij} + \sigma_\alpha^2 v_i v_j = \sigma_\alpha^2 v_i v_j. \tag{2}$$

The corresponding correlation coefficient $\tilde{\mathcal{C}}_{ij}$ is:

$$\tilde{\mathcal{C}}_{ij} = \frac{\sigma_\alpha^2 v_i v_j}{\sqrt{(1 + \sigma_\alpha^2 v_i^2)(1 + \sigma_\alpha^2 v_j^2)}}. \tag{3}$$

According to Assumption 3, there exists at least one pair of indices (m, n) with $m \neq n$ such that $v_m \neq 0$ and $v_n \neq 0$. Given $\sigma_\alpha^2 > 0$, it follows that $\tilde{\mathcal{C}}_{mn}^2 > 0$. We conclude:

$$R(\tilde{\mathbf{Z}}) = \kappa \sum_{i \neq j} \tilde{\mathcal{C}}_{ij}^2 \geq \kappa(\tilde{\mathcal{C}}_{mn}^2 + \tilde{\mathcal{C}}_{nm}^2) > 0 = R(\mathbf{Z}). \tag{4}$$

The proof is complete. □

Empirical Analysis. We present empirical results that validate the theoretical properties of $R(\mathbf{Z})$.

(i) Correlation with severity level. Our theoretical derivation establishes that the cross-dimensional correlation $\tilde{\mathcal{C}}_{ij}$ is fundamentally driven by the shift intensity variance σ_α^2. In practical terms, an increase in σ_α^2 directly corresponds to a higher severity level of the out-of-distribution (OOD) corruption. We empirically validate this direct relationship in Fig. 2. The results demonstrate a clear, positive correlation between increasing corruption severity and the redundancy score $R(\mathbf{Z})$. This confirms our theoretical hypothesis: as inputs deviate further from the source manifold (higher σ_α^2), the representation degradation exacerbates, which is precisely captured by the escalating redundancy score.

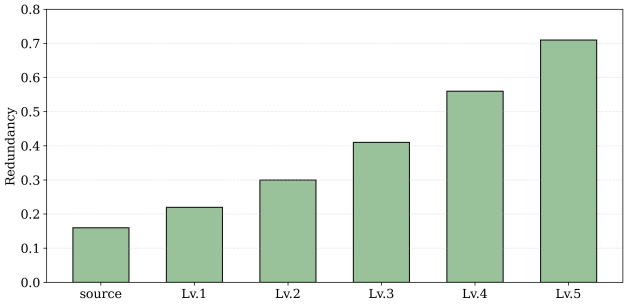

Figure 2. **Redundancy *vs*. Severity Level.** The redundancy score is evaluated across various severity levels of Kinetics50-C (Gaussian noise) and the original Kinetics50 validation set (source).

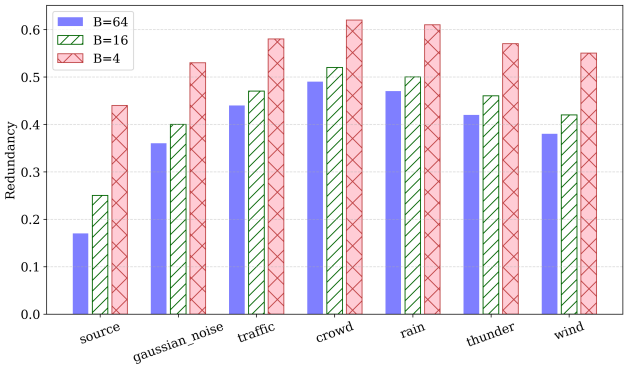

Figure 3. **Redundancy *vs*. Batch Size.** We investigate the correlation between redundancy score and batch size on VGGSound-C with audio corruptions.

(ii) Correlation with batch size. The theoretical proof relies on the population covariance matrix $\tilde{\Sigma}$ calculated over the entire distribution via expectations ($\mathbb{E}$). However, in practice, $R(\mathbf{Z})$ acts as a statistical estimator computed over a finite batch of size B. Fig. 3 assesses the stability of this sample estimator. While the absolute value of the score exhibits expected statistical variance when B is small, the results validate a critical theoretical boundary: the redundancy under distribution shift $R(\tilde{\mathbf{Z}})$ remains consistently and strictly greater than that of the source domain $R(\mathbf{Z})$, echoing the $R(\tilde{\mathbf{Z}}) > R(\mathbf{Z})$ conclusion from our theorem. To further stabilize the estimate in scenarios with small B, we suggest caching samples (*e.g.* via a momentum-based queue) to compute $R(\mathbf{Z})$ over a larger effective batch.

3. Extended Comparative Experiments

Main experiments. We report additional results for the main experiments that were not included in the main text.

Table 1. **Episodic Adaptation.** Comparison with SOTA methods on VGGSound-C with video corruptions (severity level 5) regarding **Accuracy (%, ↑)**.

Method	Noise			Blur				Weather				Digital				
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Avg.
Source	52.8	52.7	52.7	57.2	57.2	58.7	57.0	56.4	56.6	55.6	58.0	53.7	56.9	55.8	56.9	56.0
• Tent (ICLR'21)	52.7	52.7	52.7	56.7	56.7	57.9	57.1	55.9	56.3	56.3	58.4	54.0	57.4	56.2	56.7	55.8
• EATA (ICML'22)	53.0	52.8	53.0	57.0	57.0	58.1	57.2	56.3	56.8	56.8	58.7	54.1	57.6	56.4	57.0	56.1
• SAR (ICLR'23)	52.9	52.8	52.9	57.1	57.1	57.7	57.6	56.6	55.7	56.7	58.6	54.0	57.1	56.3	56.9	56.0
• READ (ICLR'24)	53.6	53.6	53.5	57.9	57.7	59.4	58.8	56.8	57.1	56.9	59.9	53.8	58.6	57.1	57.6	56.9
• TSA (ICML'25)	53.4	53.4	53.1	57.5	57.3	58.8	58.3	56.6	56.9	57.0	59.3	55.3	58.0	56.7	57.8	56.6
• **Ours**	**54.7**	**54.7**	**54.6**	**58.3**	**58.3**	**59.5**	**59.0**	**57.3**	**58.0**	**57.8**	**60.1**	**56.0**	**58.9**	**57.4**	**58.0**	**57.5**

Table 2. **Continual Adaptation.** Comparison with SOTA methods on VGGSound-C with video corruptions (severity level 5) regarding **Accuracy (%, ↑)**.

Method	Noise			Blur				Weather				Digital				
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Avg.
	$t \longrightarrow$															
Source	52.8	52.7	52.7	57.2	57.2	58.7	57.0	56.4	56.6	55.6	58.0	53.7	56.9	55.8	56.9	56.0
• Tent (ICLR'21)	52.7	52.4	51.5	54.4	54.0	55.2	54.8	52.1	52.9	52.8	52.8	50.6	52.7	52.2	51.8	52.9
• EATA (ICML'22)	53.0	53.4	53.5	56.7	57.1	58.6	58.5	55.8	56.8	57.3	58.0	55.3	57.6	57.1	57.5	56.4
• SAR (ICLR'23)	52.9	53.0	53.0	56.9	56.7	58.5	57.5	56.0	56.8	56.0	58.2	54.1	57.1	55.4	56.3	55.9
• READ (ICLR'24)	53.7	54.0	53.9	57.4	57.4	58.2	57.9	56.5	57.1	57.0	58.0	55.4	56.8	56.2	56.2	56.4
• TSA (ICML'25)	53.4	53.8	53.7	57.1	57.4	58.5	58.1	56.5	57.7	57.2	58.7	55.5	57.6	56.5	57.1	56.6
• **Ours**	**54.0**	**54.9**	**54.9**	**58.3**	**58.8**	**59.9**	**59.5**	**57.5**	**58.2**	**58.3**	**60.0**	**56.4**	**58.6**	**57.7**	**58.3**	**57.7**

Table 3. **Episodic Adaptation.** Comparison with MM-TTA methods on Kinetics50-C with video corruptions (severity level 5) regarding **Accuracy (%, ↑)**.

Method	Noise			Blur				Weather				Digital				
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	Avg.
Source	46.8	48.0	46.9	67.5	62.2	70.8	66.7	61.6	60.3	46.7	75.2	52.1	65.7	66.5	61.9	59.9
• READ (ICLR'24)	49.4	49.7	49.0	68.0	65.1	71.2	69.0	64.5	64.4	57.4	75.5	53.6	68.3	68.0	65.1	62.5
• ABPEM (AAAI'25)	50.6	51.1	50.5	68.7	66.6	**72.6**	69.6	64.4	66.2	60.5	76.0	55.3	69.5	69.2	66.2	63.8
• SuMi (ICLR'25)	50.1	50.7	50.4	68.2	65.6	72.2	69.7	65.7	67.0	56.5	**77.1**	55.2	69.3	71.2	**68.9**	63.9
• TSA (ICML'25)	50.7	51.1	50.4	67.9	67.1	71.7	69.2	65.5	66.2	61.3	75.2	56.2	69.5	68.8	66.6	63.8
• BriMPR (AAAI'26)	50.0	50.8	50.3	68.4	67.5	71.4	69.0	65.3	65.1	63.4	76.1	56.8	**71.6**	**72.2**	67.2	64.3
• **Ours**	**50.8**	**51.6**	**50.7**	**70.2**	**69.3**	72.3	**71.3**	**66.1**	**68.2**	**63.5**	75.2	**58.1**	71.2	70.5	68.6	**65.2**

These results cover the evaluation of episodic and continual adaptation on VGGSound-C with video corruptions. As the video serves as an auxiliary modality in the VGGSound dataset, the improvements observed across all methods are modest. Nonetheless, as demonstrated in Tabs. 1 and 2, our approach surpasses the current state-of-the-art methods by 0.6% and 1.1%, respectively.

Comparison with MM-TTA baselines. In the comparative experiments presented in the main text, we focus only on the READ [8] and TSA [2] methods within MM-TTA. To further demonstrate the superiority of our approach, we have included additional recently proposed baselines: ❶ **ABPEM** [9], which introduces attention bootstrapping and master entropy minimization to reduce the attention gap. ❷ **SuMi** [3] which proposes two novel strategies: sample identification with interquartile range smoothing and

unimodal assistance, and mutual information sharing. ❸ **BriMPR** [5], which employs a progressive re-alignment method to address the coupling between uni-modal and multi-modal misalignments. The experiment is performed under consistent conditions, specifically a learning rate of 1e-4 and a batch size of 64, to ensure a fair comparison. The implementation follows their official codes [5] [6] [7]. It should be noted that the BriMPR method updates the modality-specific encoders through prompt learning, which incurs significant computational costs, and also utilizes source domain statistics. However, as demonstrated in Tab. 3, our method surpasses BriMPR, the second-best approach, by 0.9%, highlighting the superior effectiveness.

References

[1] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *International Conference on Acoustics, Speech and Signal Processing*, 2020. 1

[2] Mingcai Chen, Baoming Zhang, Zongbo Han, Yuntao Du, Wenyu Jiang, Yanmeng Wang, Shuai Feng, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *International Conference on Machine Learning*, 2025. 1, 3

[3] Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *International Conference on Learning Representations*, 2025. 3

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017. 1

[5] Jiacheng Li and Songhe Feng. Bridging modalities via progressive re-alignment for multimodal test-time adaptation. In *Annual AAAI Conference on Artificial Intelligence*, 2026. 4

[6] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, 2022. 1

[7] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1

[8] Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaption against multi-modal reliability bias. In *International Conference on Learning Representations*, 2024. 1, 3

[9] Yusheng Zhao, Junyu Luo, Xiao Luo, Jinsheng Huang, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. Attention bootstrapping for multi-modal test-time adaptation. In *Annual AAAI Conference on Artificial Intelligence*, 2025. 3

[5] https://github.com/YushengZhao/ABPEM
[6] https://github.com/zrguo/SuMi
[7] https://github.com/Luchicken/BriMPR