REPRESENTATION EDITING FOR MULTIMODAL TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

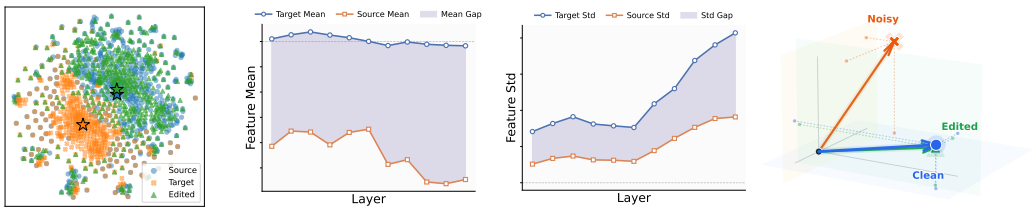
Multimodal test-time adaptation (TTA) aims to adapt a pretrained multimodal model online to distribution shift across modalities using unlabeled test data, showing broad potential in real-world applications. However, existing methods primarily focus on adjusting fused features to bridge the source-target gap, lacking explicit control over intermediate representation misalignment, which is a key driver of performance drop under distribution shift. In this work, we tackle this challenge from the perspective of representation engineering. Unlike previous TTA methods that update fusion weights in place, we propose FourIer Representation Editor (FIRE), a novel multimodal TTA approach that directly edits semantically rich intermediate representations. Specifically, we first adopt representation editors into each intermediate layer of the unimodal encoders, enabling layer-wise calibration of unimodal representations. To further enhance the diversity and stability of the low-rank editing subspaces, each representation editor performs frequency domain mixing via the fast Fourier transform to construct structured bases. Moreover, we introduce multi-level adaptation objectives to optimize these editors, jointly promoting cross-modal semantic alignment, source-target statistical alignment, and asymmetric prediction consistency. In this way, FIRE yields aligned unimodal representations for fusion and further improves prediction reliability. Extensive experiments on two widely used multimodal benchmarks under various corruption types demonstrate the superiority of our approach over existing multimodal TTA methods. The source code is available at `https://anonymous.4open.science/r/FIRE-BF81`.

1 INTRODUCTION

Multimodal learning Zheng et al. (2023); Rakib & Bagavathi (2025); Jiang et al. (2025); Xu et al. (2025); Dong et al. (2026); Hu et al. (2026) aims to integrate heterogeneous information from different sensors to improve holistic task understanding, which has received growing attention in various real-world applications. Despite its success, multimodal models often suffer from distribution shifts between source and target data in open-world scenarios due to unpredictable factors, such as weather variations and sensor failures. Consequently, the performance of static models can degrade substantially during inference Hendrycks & Dietterich (2019); Shin et al. (2022); Xu et al. (2026). To address this issue, test-time adaptation (TTA) Wang et al. (2021); Zhang et al. (2026) has emerged as a practical and effective solution: it updates model parameters online using unlabeled test data to adapt to the target distribution, transforming static inference into an adaptive process.

Most existing TTA methods, such as Tent Wang et al. (2021), EATA Niu et al. (2022), and SAR Niu et al. (2023), focus on unimodal settings and fine-tuning normalization layers via entropy minimization. However, in practice, different modalities may experience varying degrees of distribution shift in multimodal learning, making multimodal TTA substantially more challenging. Specifically, multimodal systems may encounter either corruption in a single modality or simultaneous corruption across multiple modalities. Fortunately, several studies Yang et al. (2024); He et al. (2026) have investigated multimodal distribution shifts. As a pioneering work on multimodal TTA, READ Yang et al. (2024) and its successors ABPEM Zhao et al. (2025) and TSA Chen et al. (2025), etc., address modality-specific corruption by updating the self-attention layer in the fusion module. Recent research, including SuMi Guo & Jin (2025) and PTA Wang et al. (2025), further considers more challenging scenarios where multiple modalities are corrupted simultaneously.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Representation shift. (b) Mean misalignment. (c) Std. misalignment. (d) Representation editing.

Figure 1: Overview of representation shift and editing on Kinetics50-C. (a) t-SNE visualization of multimodal representations from the source, target, and edited target domains. The star denotes the feature centroid. (b) and (c) Differences in intermediate-layer feature statistics between the source and target domains, where the mean characterizes the distribution center and the standard deviation (std) characterizes its dispersion. (d) Visualization of clean, noisy, and edited representations of the same target sample. Our method effectively mitigates the representation shift.

Although the aforementioned methods can alleviate multimodal distribution shift, they focus more on adapting fused multimodal features while lacking explicit control over the representation shifts in intermediate layers, which serve as a precursor to multimodal distribution shift. Unlike previous works, we seek a solution from the perspective of representation engineering. As shown in Figure 1 (a), we visualize multimodal features from the source and target domains on Kinetics50-C Yang et al. (2024) dataset using t-SNE Van der Maaten & Hinton (2008) and observe a clear separation between their distributions, suggesting that representation shifts are the most direct manifestation of distribution shifts. Since representation shifts typically do not emerge abruptly at the fusion stage, we believe that the misalignment in intermediate representations is a key driver of performance drop under distribution shift. Therefore, we quantify the statistical discrepancy between source and target features in intermediate layers. As shown in Figure 1 (b) and (c), the results indicate that the shift emerges in shallow layers and tends to increase with depth. This naturally raises the question: can we directly modify intermediate representations to align the target with the source domain, thereby enhancing the quality of fusion-layer representation and improving prediction reliability?

Motivated by the above insights, we propose a novel Multimodal TTA approach, FourIer Representation Editor (FIRE), to directly edit semantically rich intermediate representations, facilitating alignment between the source and target domains. Specifically, we first introduce representation editors and place them after each transformer layer in every unimodal encoder, jointly optimizing unimodal representations layer-wise. By layer-wise calibrating unimodal representations, editors align the target distribution with the source, thereby facilitating reliable fusion. Notably, only the editors are updated in TTA, while the pretrained model remains entirely frozen. Subsequently, we enhance the representation editor with fast Fourier transforms Reddy (1998); Chi et al. (2020). By incorporating frequency domain information into the low-rank subspace, FIRE integrates global coupling across dimensions, enriching the editing directions and improving adaptation stability. Furthermore, we optimize these editors with multi-level adaptation objectives, including cross-channel semantic alignment, source-target statistical calibration, and asymmetric prediction consistency. To this end, FIRE effectively calibrates intermediate representations and provides aligned representations for fusion, thereby improving prediction reliability. As shown in Figure 1(a) and (d), FIRE significantly realigns the representation distribution. Extensive experiments on widely used multimodal TTA benchmarks demonstrate that FIRE achieves significant performance improvements.

2 Related Work

2.1 Multimodal Test-Time Adaptation

Test-time adaptation (TTA) Li et al. (2025); Wu et al. (2025); Jia et al. (2025); Fan et al. (2026); Wu et al. (2026) aims to online adapt a model to the current target distribution using unlabeled test samples. Most TTA methods mainly focus on unimodal learning settings Li et al. (2024); Zhang et al. (2025); Chu et al. (2026); Yang et al. (2026). A representative method, TENT Wang et al. (2021), updates the normalization layers by minimizing prediction entropy. Subsequent works, including

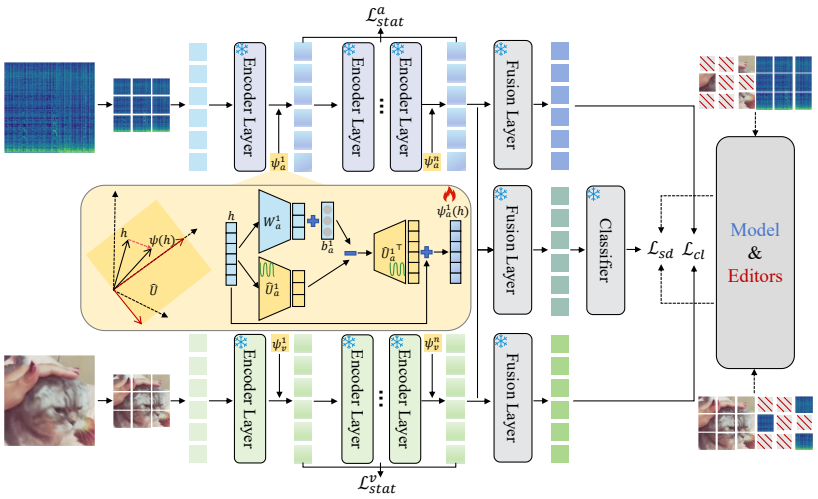


Figure 2: Overview of FIRE. The entire model remains frozen, with only editors being updated. The editor aims to identify an editing vector that modifies the representations only within the linear subspace spanned by the rows of $\hat{U}$. "Model & Editors" is the same pretrained model and editors. In particular, we illustrate the representation editing process in a 3-dimensional space.

EATA Niu et al. (2022), SAR Niu et al. (2023), and DeYO Lee et al. (2024), design sample filtering strategies to promote adaptation with more informative target samples.

Multimodal TTA is more challenging, as different modalities may suffer from corruptions of varying severity. As a pioneering work on multimodal TTA, READ Yang et al. (2024) highlights the reliability bias caused by unimodal corruption and addresses it by optimizing a self-attention Vaswani et al. (2017) fusion layer. Its successors further improve multimodal TTA with strategies such as attention bootstrapping Zhao et al. (2025), smoothing the shift Guo & Jin (2025), and partition-then-adapt Wang et al. (2025). Other works Chen et al. (2025); Li & Feng (2026) begin to focus on unimodal information, aiming to bridge the source-target gap by correcting unimodal shifts before fusion. However, these methods lack explicit control over intermediate representation shifts. In this paper, we aim to address distribution shifts by editing unimodal representations in a layer-wise manner.

2.2 Representation Engineering

Representation engineering Zou et al. (2023); Subramani et al. (2022); Burns et al. (2023); Bartoszcze et al. (2025) aims to understand and control model behavior from the perspective of internal neural representations. Recent studies have explored the potential of representation-level model adaptation. Specifically, RED Wu et al. (2024a) performs direct representation editing by applying element-wise scaling and shifting to the full hidden representations in LLMs Touvron et al. (2023). ReFT Wu et al. (2024b) further controls partial representations within specific subspaces using low-rank projection matrices with orthogonal rows and linear projectors. These methods show that directly editing internal representations can effectively regulate model behavior. Extending from unimodal models to multimodal settings, recent works Liu et al. (2025); Das & Jena (2026) begin to investigate representation editing in multimodal models. MRT Liu et al. (2025) transfers representation editors from LLMs to MLLMs Liu et al. (2023). However, these methods are mainly designed for LLMs or MLLMs parameter-efficient fine-tuning (PEFT) Houlsby et al. (2019); Hu et al. (2022), and their effectiveness in multimodal TTA remains largely unexplored. Therefore, our approach fills this gap by layer-wise editing intermediate representations at test time, enabling representation-level adaptation for multimodal TTA.

3 Methodology

In this section, we introduce FIRE, a novel multimodal TTA method that effectively aligns the target and source distributions. Overview of our approach is provided in Figure 2.

Algorithm 1 Algorithm of Fourier Representation Editor

Input: The intermediate representation $\boldsymbol{h}$ and editor parameters $\{\boldsymbol{U}, \boldsymbol{W}, \boldsymbol{b}\}$.
Output: The edited representation.

1: Compute the FFT of $\boldsymbol{U}$ using Equation 3;
2: Compute the enhanced subspace $\hat{\boldsymbol{U}}$ using Equation 4;
3: Compute the subspace representation $\boldsymbol{h}' = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$;
4: Compute the subspace correction $\Delta\boldsymbol{h} = \hat{\boldsymbol{U}}^\top(\boldsymbol{h}' - \hat{\boldsymbol{U}}\boldsymbol{h})$;
5: Obtain edited representation $\psi(\boldsymbol{h}) = \boldsymbol{h} + \Delta\boldsymbol{h}$;
6: **Return** representation $\psi(\boldsymbol{h})$.

3.1 Problem Statement

Without loss of generality, we take two modalities, i.e., audio and video, as a showcase for clarity of presentation. For the sake of simplicity, we use superscript r to indicate the module corresponding to a specific modality in this section, where $r \in \{a, v\}$. For clarity, we use $F_\Theta = \{f_{\Theta^a}, f_{\Theta^v}, f_{\Theta^m}, f_{\Theta^c}\}$ to denote the off-the-shelf model that trained on the source domain $\mathcal{D}_\mathcal{S} = \{(\boldsymbol{x}_i^a, \boldsymbol{x}_i^v, y_i)\}_{i=1}^{N_\mathcal{S}}$, where f_{Θ^a} and f_{Θ^v} are the specific encoders for modality a and v, f_{Θ^m} and f_{Θ^c} are the fusion layer and the following classifier head. Assume that the target domain $\mathcal{D}_\mathcal{T} = \{(\boldsymbol{x}_j^a, \boldsymbol{x}_j^v)\}_{j=1}^{N_\mathcal{T}}$, where the target data follow the distribution $P_\mathcal{T}(\boldsymbol{x})$. During training, the model F_Θ would fit the distribution of source data, i.e., $P_\mathcal{S}(\boldsymbol{x})$. As a result, in the inference stage, when a domain shift exists between the source and target distributions, i.e., $P_\mathcal{S}(\boldsymbol{x}) \neq P_\mathcal{T}(\boldsymbol{x})$, the performance of F_Θ deteriorates significantly. The goal of TTA is to adapt F_Θ online to the target domain.

3.2 Fourier Representation Editor

Since multiple studies Smolensky (1986); Lasri et al. (2022); Guerner et al. (2023) have shown that human-interpretable concepts are encoded linearly and that linear subspace interventions can correspond to specific semantic attributes, we facilitate representation editing by targeting linear subspaces. The final prediction of a multimodal model is determined by the combined contributions of multiple individual modalities, which naturally enables multimodal TTA to be formulated as a set of unimodal TTA problems. We apply representation editors after each transformer layer of the unimodal encoders and jointly optimize all unimodal representations. Notably, during adaptation, only these editors are updated, while the entire model remains completely frozen.

Representation Editor. Inspired by the linear representation hypothesis Wu et al. (2024b), we introduce a representation editor to edit semantically rich intermediate representations. The editor updates the original representation $\boldsymbol{h}$ within a specific subspace to reflect the desired intervention obtained from a linear projection $\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$, where both $\boldsymbol{W}$ and $\boldsymbol{b}$ are learnable parameters. Then, the editing operation is confined to the subspace spanned by the rows of a low-rank matrix $\boldsymbol{U}$ with orthonormal rows. We employ QR decomposition to maintain the row orthonormality of $\boldsymbol{U}$. Therefore, the editor only adjusts the aspects of the representation associated with the target attributes while preserving the remaining semantic information. The editor can be formed as:

$$\Psi(\boldsymbol{h}) = \boldsymbol{h} + \boldsymbol{U}^\top(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} - \boldsymbol{U}\boldsymbol{h}), \tag{1}$$

where $\boldsymbol{U} \in \mathbb{R}^{d_m \times d_n}$ and $\boldsymbol{W} \in \mathbb{R}^{d_m \times d_n}$ are low-rank matrices. d_m is the rank of the subspace and d_n is the hidden dimensionality, i.e., $d_m \ll d_n$. Specifically, $\boldsymbol{U}\boldsymbol{h}$ projects the original representation $\boldsymbol{h}$ onto the subspace spanned by $\boldsymbol{U}$, while $\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$ provides the target values within that subspace via a linear projection from $\boldsymbol{h}$. The difference $\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} - \boldsymbol{U}\boldsymbol{h}$ computes the necessary intervention, which is subsequently mapped back into the original space via $\boldsymbol{U}^\top$. By adding this intervention to $\boldsymbol{h}$, we obtain the edited representation $\Psi(\boldsymbol{h})$ that incorporates the desired modifications while maintaining the components orthogonal to the subspace $\boldsymbol{U}$.

Fourier-Enhanced Subspace. The FFT is a powerful algorithm for computing the Discrete Fourier Transform (DFT). Given a sequence $\{\boldsymbol{z}_m\}$ where m is a member of the interval $m \in [0, M-1]$, the

DFT is defined as:

$$\mathcal{F}(\boldsymbol{z}) = \mathcal{Z}_k = \sum_{m=0}^{M-1} \boldsymbol{z}_m e^{-i2\pi \frac{k}{M} m}, 0 \leq k \leq M-1. \tag{2}$$

Recent parameter-efficient fine-tuning studies Gao et al. (2024); Zeng et al. (2024); Borse et al. (2024) demonstrate that Fourier transforms can effectively bridge the gap between pretraining and fine-tuning datasets by incorporating additional information. Motivated by this, we explore integrating FFT operations into the representation editor to support on-the-fly adjustments during TTA. To this end, we introduce the FFT to impose global coupling across dimensions in the low-rank matrix $\boldsymbol{U}$, thereby enriching editable directions and enabling more stable adaptation of the representation editor. Specifically, we first apply an FFT to the low-rank matrix $\boldsymbol{U}$ along the hidden dimensions:

$$\boldsymbol{U}_{\mathcal{F}} = \Re(\mathcal{F}(\boldsymbol{U})). \tag{3}$$

To maintain consistency with the original pretrained representation space, we use the FFT only to construct the editing basis and retain its real component $\Re$, ensuring that the final editing basis and representation updates remain in the real-valued space. We then enhance the subspace by combining the editing directions in the original space with the Fourier-domain editing directions, formulated as:

$$\hat{\boldsymbol{U}} = \mathrm{qf}((\boldsymbol{U} + \boldsymbol{U}_{\mathcal{F}})^\top)^\top, \tag{4}$$

where $\mathrm{qf}(\cdot)$ denotes the orthogonal factor obtained from the QR decomposition.

To improve efficiency, we adopt the Cooley-Tukey FFT algorithm Cooley & Tukey (1965), following common practice Lee-Thorp et al. (2022). The final Fourier representation editor is defined as:

$$\psi(\boldsymbol{h}) = \boldsymbol{h} + \hat{\boldsymbol{U}}^\top(\boldsymbol{W}\boldsymbol{h} + \boldsymbol{b} - \hat{\boldsymbol{U}}\boldsymbol{h}). \tag{5}$$

Therefore, the representation editing operation is constrained to the Fourier-enhanced low-rank orthogonal subspace, enabling the method to introduce more diverse editing directions and improve stability without adding extra parameters. To ensure stable TTA, we initialize the representation editor in a no-op scheme, unlike common methods Wu et al. (2024b); Liu et al. (2025). Our algorithm is summarized in Algorithm 1. Moreover, we provide a theoretical analysis of the stability and effectiveness of the self-training subspace in the appendix.

3.3 Adaptation through Self-Training

Due to the absence of supervision at test time, we optimize the representation editors with self-training objectives. Inspired by related tasks Niu et al. (2024); Radford et al. (2021); Ramazanova et al. (2025); Li & Feng (2026), we constrain editor adaptation at multiple levels, including cross-modal semantic alignment, source-target statistical alignment, and asymmetric prediction consistency.

Cross-modal Semantic Alignment Loss. The cross-modal semantic alignment loss aims to realign semantic relationships across modalities. Cross-modal contrastive learning Radford et al. (2021) is a key paradigm for multimodal learning and can enhance semantic alignment across modalities. In multimodal TTA, we introduce contrastive loss to help the model realign the semantics of shifted and unshifted modalities. Specifically, given final unimodal representations $\tilde{\boldsymbol{h}}_a$ and $\tilde{\boldsymbol{h}}_v$, the cross-modal semantic alignment loss is defined as:

$$\begin{aligned}\mathcal{L}_{cl} = -\frac{1}{2B}\sum_{j=1}^{B}(&\log\frac{e^{sim(\tilde{\boldsymbol{h}}_{a,j},\tilde{\boldsymbol{h}}_{v,j})/\tau}}{\sum_{k=1}^{B} e^{sim(\tilde{\boldsymbol{h}}_{a,j},\tilde{\boldsymbol{h}}_{v,k})/\tau}} \\ &+ \log\frac{e^{sim(\tilde{\boldsymbol{h}}_{a,j},\tilde{\boldsymbol{h}}_{v,j})/\tau}}{\sum_{k=1}^{B} e^{sim(\tilde{\boldsymbol{h}}_{a,k},\tilde{\boldsymbol{h}}_{v,j})/\tau}}),\end{aligned} \tag{6}$$

where B denotes the mini-batch size, $sim(\cdot,\cdot)$ denotes the cosine similarity, and τ denotes the temperature hyperparameter set to its default value..

Source-target Statistical Alignment Loss. Source-target statistical alignment loss is an effective distribution calibration objective, which calibrates target representations by aligning their layer-wise statistics with those of the source domain. Specifically, we offline compute the mean $\{\mu_l^{\mathcal{S},r}\}_{l=1}^{n}$

Table 1: Comparison with SOTA methods on Kinetics50-C (top) and VGGSound-C (bottom) benchmarks with corrupted video modality (severity level 5).

	Noise			Blur				Weather				Digital				
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elast.	Pixel.	Jpeg	Avg.
Source	48.2	50.0	49.2	67.7	61.6	70.6	66.1	60.9	60.7	44.7	75.9	51.8	65.5	68.7	66.1	60.5
• $\text{Tent}_{\text{ICLR21}}$	48.2	49.8	48.7	67.7	62.1	70.8	67.2	61.8	61.4	33.7	76.0	51.2	66.6	69.6	66.9	60.1
• $\text{T3A}_{\text{NeurIPS21}}$	47.8	49.0	48.2	62.8	59.3	65.5	62.0	60.2	57.5	50.7	69.8	47.6	62.8	65.7	64.5	58.2
• $\text{EATA}_{\text{ICML22}}$	48.7	50.4	49.6	67.8	63.2	70.8	67.5	62.5	62.5	47.9	76.1	52.2	66.9	69.7	67.4	61.5
• $\text{SAR}_{\text{ICLR23}}$	48.5	50.2	49.2	67.8	63.8	70.9	67.9	63.1	62.7	38.7	76.1	52.2	67.1	69.8	67.4	61.0
• $\text{DeYO}_{\text{ICLR24}}$	48.6	50.2	49.4	67.9	62.6	70.9	67.4	62.5	62.3	40.4	76.1	52.2	66.8	69.8	67.3	61.0
• $\text{FOA}_{\text{ICML24}}$	49.2	50.8	49.7	66.0	65.5	69.8	67.4	62.8	65.7	60.3	74.9	51.9	69.5	68.8	68.0	62.7
• $\text{READ}_{\text{ICLR24}}$	50.7	52.2	51.4	67.9	65.3	71.1	68.7	64.0	65.8	56.3	76.3	53.6	68.7	70.0	68.6	63.4
• $\text{ABPEM}_{\text{AAAI25}}$	52.1	53.1	52.8	69.0	65.6	71.8	68.8	64.1	65.7	57.9	76.6	54.3	69.2	71.1	69.2	64.1
• $\text{SuMi}_{\text{ICLR25}}$	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
• $\text{TSA}_{\text{ICML25}}$	52.6	52.3	52.0	68.7	68.0	70.7	68.8	65.2	66.6	64.3	74.6	57.4	70.5	69.0	66.2	64.5
• $\text{PTA}_{\text{NeurIPS25}}$	52.9	53.0	52.5	68.5	66.9	72.4	69.1	64.9	67.1	64.1	75.9	55.1	69.2	70.2	68.4	64.7
• $\text{BriMPR}_{\text{AAAI26}}$	55.3	56.1	56.7	67.8	67.9	70.6	68.8	65.9	66.2	64.1	76.2	56.3	72.0	73.7	70.5	65.9
• FIRE	**59.3**	**59.8**	**59.9**	**70.0**	**71.1**	**73.4**	**71.8**	**68.7**	**69.3**	**67.3**	76.6	**61.4**	**74.2**	**75.1**	**71.8**	**68.6**
Source	52.9	53.0	53.1	57.2	57.2	58.5	57.5	56.5	57.1	55.6	59.2	53.7	57.1	56.4	57.3	56.2
• $\text{Tent}_{\text{ICLR21}}$	53.2	53.3	53.3	56.8	56.6	57.9	57.2	55.9	56.6	56.5	58.5	53.9	57.5	56.8	56.9	56.1
• $\text{T3A}_{\text{NeurIPS21}}$	53.2	53.1	53.3	57.2	57.3	58.5	58.0	56.6	57.2	56.7	59.4	54.3	57.8	56.9	57.5	56.5
• $\text{EATA}_{\text{ICML22}}$	53.4	53.5	53.5	57.0	57.0	58.3	57.7	56.3	57.0	56.8	59.1	54.2	57.9	57.2	57.2	56.4
• $\text{SAR}_{\text{ICLR23}}$	53.3	53.3	53.3	56.4	56.5	57.9	57.3	55.6	56.4	56.3	58.8	53.7	57.8	56.9	57.0	56.0
• $\text{DeYO}_{\text{ICLR24}}$	53.3	53.4	53.4	56.7	56.7	58.0	57.3	56.0	56.8	56.4	58.7	53.9	57.7	57.0	57.0	56.2
• $\text{FOA}_{\text{ICML24}}$	52.7	52.7	52.7	53.2	53.6	53.6	53.8	53.4	53.4	53.3	55.6	52.5	55.3	53.7	54.4	53.6
• $\text{READ}_{\text{ICLR24}}$	53.8	54.0	53.8	58.0	57.9	59.2	58.7	57.1	**58.2**	50.0	60.0	55.2	58.5	57.7	58.2	56.7
• $\text{ABPEM}_{\text{AAAI25}}$	46.5	46.7	46.5	54.2	55.1	56.4	55.2	51.3	53.2	52.1	56.6	52.1	54.4	51.7	54.7	52.4
• $\text{SuMi}_{\text{ICLR25}}$	54.0	54.3	53.8	58.2	58.4	59.4	58.7	57.5	**58.2**	57.6	59.4	54.8	59.0	57.5	58.2	57.3
• $\text{TSA}_{\text{ICML25}}$	53.9	53.9	54.0	57.6	58.0	59.0	58.6	56.9	57.0	56.6	59.8	54.7	58.6	56.7	57.9	56.9
• $\text{PTA}_{\text{NeurIPS25}}$	48.3	48.2	48.1	55.9	56.5	57.8	56.4	53.9	54.2	53.2	58.3	52.0	55.8	53.6	56.3	53.9
• $\text{BriMPR}_{\text{AAAI26}}$	54.9	55.0	55.0	57.9	58.5	58.9	58.7	57.5	58.0	58.5	60.3	54.5	59.7	59.3	59.0	57.7
• FIRE	**55.5**	**55.5**	**55.5**	**58.4**	**59.0**	**59.6**	**59.3**	**57.8**	**58.2**	**58.7**	**60.5**	**56.2**	**60.1**	**59.7**	**59.4**	**58.2**

and standard deviation $\{\sigma_l^{\mathcal{S},r}\}_{l=1}^n$ of the source-domain unimodal representations at each layer after model pretraining and before TTA, where l denotes the layer index. During TTA, given a mini-batch of test samples, we obtain the edited representation sequence from the l-th layer of a specific modality encoder. We then compute the corresponding unimodal representation statistics of the test samples, $\{\mu_l^{\mathcal{T},r}\}_{l=1}^n$ and $\{\sigma_l^{\mathcal{T},r}\}_{l=1}^n$, and align them with the precomputed source statistics. This enables target representation calibration toward the source distribution, requiring only cached source statistics rather than access to the source domain data. Formally, the loss is defined by:

$$\begin{aligned}\mathcal{L}_{stat}(\mu^{\mathcal{S}},\sigma^{\mathcal{S}},\mu^{\mathcal{T}},\sigma^{\mathcal{T}}) &= \sum_{r\in\{a,v\}} \mathcal{L}_{stat}^{r}(\mu^{\mathcal{S}},\sigma^{\mathcal{S}},\mu^{\mathcal{T}},\sigma^{\mathcal{T}}) \\ &= \sum_{r\in\{a,v\}} \frac{1}{n}\sum_{l=1}^{n}(\|\mu_l^{\mathcal{T},r}-\mu_l^{\mathcal{S},r}\|_2 + \|\sigma_l^{\mathcal{T},r}-\sigma_l^{\mathcal{S},r}\|_2),\end{aligned} \tag{7}$$

where $\|\cdot\|_2$ denotes the Euclidean norm. Thus, statistics matching provides a lightweight proxy for source-target representation alignment without assuming a specific representation distribution.

Asymmetric Prediction Consistency Loss. To improve model predictions, we introduce an asymmetric prediction consistency loss. Ideally, the model F_Θ should produce consistent predictions under both complete and incomplete modalities, making it robust to distribution shifts. Specifically, for each test sample, we randomly mask a portion of information from one modality (75% by default) while keeping the other modality unchanged, yielding two sets of incomplete sample pairs. These two sets are then fed into the model and layer-wise calibrated by the representation editor at each layer to obtain the corresponding predictions. Finally, we use self-distillation to encourage the model to be insensitive to the specific modality source present during testing. The loss is formulated as:

$$\mathcal{L}_{CE}^{r}(\boldsymbol{p}^r,\boldsymbol{p}) = \frac{1}{B}\sum_{j=1}^{B}\ell(\boldsymbol{p}_j^r,\boldsymbol{p}_j) = -\frac{1}{B}\sum_{j=1}^{B}\boldsymbol{p}_j^{\top}\log(\boldsymbol{p}_j^r), \tag{8}$$

where $\boldsymbol{p}^r$ denotes the prediction after masking modality r, $\boldsymbol{p}$ denotes the prediction with complete modalities, and ℓ is the cross entropy loss. Soft predictions as consistency targets avoid committing to potentially incorrect hard pseudo-labels, thereby reducing the risk of confirmation bias.

Table 2: Comparison with SOTA methods on Kinetics50-C (left) and VGGSound-C (right) benchmarks with corrupted audio modality (severity level 5).

	Noise			Weather				Noise			Weather			
Method	Gauss.	Traff.	Crowd	Rain	Thund.	Wind	Avg.	Gauss.	Traff.	Crowd	Rain	Thund.	Wind	Avg.
Source	74.3	65.3	68.0	70.3	68.0	70.5	69.4	37.3	21.2	16.9	21.8	27.3	25.7	25.0
• $\text{Tent}_{\text{ICLR21}}$	74.6	67.4	69.5	70.8	67.6	71.2	70.2	10.8	2.8	1.8	2.9	5.6	3.9	4.6
• $\text{T3A}_{\text{NeurIPS21}}$	71.4	63.0	65.4	66.8	66.1	67.8	66.8	39.8	28.2	24.8	28.6	34.7	30.6	31.1
• $\text{EATA}_{\text{ICML22}}$	74.6	67.3	69.4	70.8	69.8	71.0	70.5	40.2	30.0	27.8	29.7	36.5	32.2	32.7
• $\text{SAR}_{\text{ICLR23}}$	74.6	67.0	69.2	70.9	69.5	70.9	70.3	30.4	5.5	8.0	9.3	32.5	17.2	17.1
• $\text{DeYO}_{\text{ICLR24}}$	74.6	67.0	69.3	70.8	69.0	71.0	70.3	22.9	4.9	15.8	4.9	16.5	20.0	14.2
• $\text{FOA}_{\text{ICML24}}$	73.8	70.0	70.5	71.0	73.0	71.2	71.6	31.5	26.2	23.7	31.0	34.2	26.7	28.9
• $\text{READ}_{\text{ICLR24}}$	74.8	69.2	69.9	71.4	72.4	71.0	71.5	39.9	29.4	26.8	30.8	36.8	30.7	32.4
• $\text{ABPEM}_{\text{AAAI25}}$	74.7	68.5	70.3	71.7	72.3	71.2	71.4	38.5	27.6	25.2	26.5	32.7	26.5	29.5
• $\text{SuMi}_{\text{ICLR25}}$	**75.1**	68.9	70.6	71.6	72.8	**72.1**	71.9	**41.9**	26.3	27.9	31.6	37.1	34.1	33.2
• $\text{TSA}_{\text{ICML25}}$	74.5	69.6	70.5	71.4	72.0	71.0	71.5	41.5	31.8	30.9	**32.6**	38.9	32.6	34.7
• $\text{PTA}_{\text{NeurIPS25}}$	73.5	70.0	70.4	70.3	73.4	70.8	71.4	40.5	30.8	31.8	32.0	38.6	31.7	34.2
• $\text{BriMPR}_{\text{AAAI26}}$	74.8	69.6	71.7	71.5	72.4	72.0	72.0	39.3	35.0	36.7	32.5	41.0	34.6	36.5
• FIRE	74.8	**72.4**	**73.5**	**72.4**	**75.2**	**72.1**	**73.4**	39.9	**35.5**	**37.2**	32.4	**41.6**	**35.1**	**36.9**

Table 3: Comparison with SOTA methods on Kinetics50-C benchmarks for multimodal domain shifts (severity level 5).

	Noise			Blur				Weather				Digital				
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elast.	Pixel.	Jpeg	Avg.
Source	13.1	14.1	13.3	37.2	37.4	45.3	41.8	29.4	32.6	20.4	55.2	18.3	42.5	38.8	37.8	31.8
• $\text{Tent}_{\text{ICLR21}}$	9.1	9.7	9.1	32.5	34.1	43.5	40.2	23.2	28.3	13.2	55.1	13.7	40.7	34.7	35.0	28.1
• $\text{T3A}_{\text{NeurIPS21}}$	12.3	13.8	12.7	34.7	35.5	42.0	39.3	29.2	28.7	20.9	51.9	17.1	40.6	37.5	37.4	30.2
• $\text{EATA}_{\text{ICML22}}$	12.9	14.0	13.1	38.1	38.7	46.9	43.1	30.6	33.0	20.2	56.5	18.2	43.7	40.7	39.0	32.6
• $\text{SAR}_{\text{ICLR23}}$	11.8	12.8	11.9	37.4	38.2	46.3	43.1	29.8	33.0	17.4	56.0	16.0	43.5	39.5	38.2	31.7
• $\text{DeYO}_{\text{ICLR24}}$	11.0	12.0	11.1	37.0	37.7	46.3	43.2	29.9	33.3	17.9	56.2	17.0	43.7	39.7	38.0	31.6
• $\text{FOA}_{\text{ICML24}}$	18.7	20.6	19.3	43.7	45.5	50.2	47.9	38.9	43.7	37.2	60.5	23.5	52.7	48.9	47.4	39.9
• $\text{READ}_{\text{ICLR24}}$	14.5	14.9	14.8	43.8	42.1	51.0	46.5	35.4	38.9	27.6	58.9	22.6	47.1	42.1	38.1	35.9
• $\text{ABPEM}_{\text{AAAI25}}$	19.2	20.7	19.7	46.2	44.2	51.9	47.9	38.1	41.1	32.6	59.9	25.3	49.4	48.8	45.6	39.4
• $\text{SuMi}_{\text{ICLR25}}$	12.5	13.6	12.6	37.0	37.9	45.9	42.3	29.3	32.7	19.7	55.7	17.8	42.7	38.3	36.9	31.7
• $\text{TSA}_{\text{ICML25}}$	18.1	18.7	17.9	47.1	47.7	53.0	50.1	41.0	45.1	41.6	59.6	29.5	53.1	48.4	45.2	41.1
• $\text{PTA}_{\text{NeurIPS25}}$	21.9	23.0	22.1	47.7	45.9	52.6	49.3	40.3	43.6	39.7	60.0	27.3	50.4	50.9	47.6	41.5
• $\text{BriMPR}_{\text{AAAI26}}$	22.9	24.2	24.1	43.6	45.4	49.5	48.2	38.0	40.8	36.8	59.8	27.1	52.8	52.7	47.9	40.9
• FIRE	**30.2**	**30.8**	**31.4**	**48.7**	**52.2**	**54.2**	**53.3**	**44.5**	**47.5**	**45.7**	**61.9**	**35.1**	**58.2**	**58.5**	**53.2**	**47.0**

In multimodal TTA, different modalities contribute unequally to adaptation Yang et al. (2024). Treating all incomplete sample pairs equally may degrade unshifted modalities. Therefore, when one modality is shifted, we encourage masking the other unshifted modality, forcing the representation editor to learn target-to-source alignment. Based on this insight, asymmetric prediction consistency loss weights the losses of all incomplete sample pairs according to the discrepancy between the current batch statistics of each modality and the corresponding source domain statistics:

$$\mathcal{L}_{pc}(\boldsymbol{p}^a, \boldsymbol{p}^v, \boldsymbol{p}) = \frac{\mathcal{L}^v_{stat}}{\mathcal{L}_{stat}} \mathcal{L}^a_{CE}(\boldsymbol{p}^a, \boldsymbol{p}) + \frac{\mathcal{L}^a_{stat}}{\mathcal{L}_{stat}} \mathcal{L}^v_{CE}(\boldsymbol{p}^v, \boldsymbol{p}). \tag{9}$$

Multi-level Adaptation Objective. The final multi-level adaptation objective combines these three losses:

$$\mathcal{L}_{all} = \mathcal{L}_{cl} + \mathcal{L}_{stat} + \mathcal{L}_{pc}. \tag{10}$$

By minimizing this final combined loss, we can optimize the representation editors to achieve target-to-source alignment at test time.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate FIRE on the widely used multimodal datasets Kinetics50-C and VGGSound-C Yang et al. (2024), which are corrupted versions of Kinetics50 Kay et al. (2017) and VGGSound Chen et al. (2020) datasets, respectively. Specifically, Kinetics50 contains YouTube videos of human

actions across 50 categories, while VGGSound contains 309 audio-visual classes covering a wide range of real-world audio-visual events. Following Yang et al. (2024), we introduce 6 and 15 types of corruptions for the audio and video modalities, respectively, with each corruption type divided into 5 severity levels to simulate real-world multimodal distribution shifts. For multimodal corruptions, we use combinations of unimodal corruptions, resulting in 90 combinations for each benchmark. In particular, following Yang et al. (2024); Chen et al. (2025); Wang et al. (2025); Li & Feng (2026), we select the highest corruption severity level. More details are provided in the appendix.

Baselines. We select two categories of methods for comparison. In detail, unimodal TTA methods: Tent Wang et al. (2021), T3A Iwasawa & Matsuo (2021), EATA Niu et al. (2022), SAR Niu et al. (2023), DeYO Lee et al. (2024), and FOA Niu et al. (2024). Multimodal TTA methods: READ Yang et al. (2024), ABPEM Zhao et al. (2025), SuMi Guo & Jin (2025), TSA Chen et al. (2025), PTA Wang et al. (2025), and BriMPR Li & Feng (2026).

Table 4: Comparison with TTA baseline methods on VGGSound-C for multimodal domain shifts (severity level 5).

	Noise			Weather			
Method	Gauss.	Traff.	Crowd	Rain	Thund.	Wind	Avg.
Source	17.1	6.4	5.4	6.0	13.5	8.8	9.5
• Tent	3.2	0.9	0.8	0.9	2.8	1.3	1.6
• T3A	20.6	9.2	7.8	8.6	17.4	10.9	12.4
• EATA	21.5	7.7	7.1	7.3	17.3	11.9	12.1
• SAR	10.7	1.8	1.6	2.3	12.8	3.1	5.4
• DeYO	6.7	1.2	1.3	1.3	9.3	2.9	3.8
• FOA	18.8	10.8	11.4	11.6	20.5	10.4	13.9
• READ	20.1	12.5	10.7	10.5	20.5	13.4	14.6
• ABPEM	21.9	13.4	12.3	10.9	20.4	12.4	15.2
• SuMi	17.0	6.8	5.7	6.2	13.4	8.8	9.7
• TSA	20.5	12.6	12.2	12.2	22.0	13.3	15.5
• PTA	<u>24.3</u>	15.3	15.3	13.7	23.8	15.0	17.9
• BriMPR	23.5	<u>18.8</u>	<u>21.4</u>	<u>15.8</u>	<u>26.8</u>	<u>18.3</u>	<u>20.7</u>
• FIRE	**25.8**	**19.6**	**22.4**	**16.9**	**27.5**	**19.7**	**22.0**

Implementation Details. Following prior work Yang et al. (2024), we use the pretrained CAV-MAE Gong et al. (2023) as the source model. For all experiments, we set the learning rate to 1e-4 and employ the Adam Kingma & Ba (2015) optimizer. The batch size is set to 16 for Kinetics50-C and 64 for VGGSound-C. The subspace $\hat{\boldsymbol{U}}$ is orthogonally reparameterized to preserve its orthogonality during TTA. The matrix $\boldsymbol{W}$ is initialized with the initial orthogonal basis $\hat{\boldsymbol{U}}$, while $\boldsymbol{b}$ is initialized to 0, yielding a no-op initialization to ensure stability. Unless otherwise specified, the rank of the low-rank matrices is set to 6. Notably, larger rank can yield better performance, while increasing the parameter count. For competing methods, we adopt the hyperparameters recommended in respective papers. All experiments are conducted on NVIDIA RTX-3090 GPUs with 5 random seeds. Following prior work Yang et al. (2024), we report the mean results over five random seeds.

4.2 Main Results

We report the results under unimodal and multimodal shift settings in Table 1, Table 2, Table 3, and Table 4, where **bold** and <u>underlined</u> denote the best and second-best results, respectively. All results are averaged over 5 random seeds.

Robustness to unimodal shift setting. In Table 1 and Table 2, we evaluate the unimodal shift setting on Kinetics50-C and VGGSound-C datasets, where the audio and video modalities are corrupted, respectively. We can observe that: (1) Unimodal TTA methods exhibit limited effectiveness in the multimodal setting and may even suffer from negative transfer, whereas multimodal TTA methods yield consistent improvements. (2) Our approach outperforms SOTA baselines, achieving the best average performance across datasets.

Robustness to multimodal shift setting. Table 3 and Table 4 present results under the more challenging multimodal shift setting, where online data from both audio and video modalities is corrupted. Each entry in the table denotes the average performance over all corruptions of the other modality under a specific corruption of one modality. We can draw conclusions: (1) Most TTA methods yield limited gains and may even suffer from negative transfer. (2) Our approach achieves substantial improvements in this more complex setting.

4.3 Further Analysis

Impact of Rank. In the representation editor, the number of ranks directly determines the number of trainable parameters and editing directions. To further investigate the impact of rank on TTA perfor-

Table 5: Impact of rank (left) and editing depth (right) on Kinetics50-C dataset.

Rank	audio	video	both	Depth	audio	video	both
2	72.8	67.3	44.5	First Layer	71.3	63.3	36.9
4	73.3	68.2	46.3	First Half	72.7	67.7	45.2
6	73.4	68.6	47.0	Latter Half	72.7	65.7	42.2
8	**73.5**	69.0	47.5	Last Layer	71.2	63.3	36.5
10	**73.5**	**69.2**	**48.0**	All Layers	**73.4**	**68.6**	**47.0**

Table 6: The number of trainable parameters and performance on Kinetics50-C dataset.

Method	layer	audio	video	both	Params (M)
Tent	Normal	70.2	60.1	28.1	0.218
READ	Attention	71.5	63.4	35.9	1.772
TSA	Adapter	71.5	64.5	41.1	1.180
BriMPR	Prompt	72.0	65.9	40.9	0.169
FIRE (Rank2)	Editor	**72.8**	**67.3**	**44.5**	**0.068**

mance, we conduct a comprehensive study of the editor rank on Kinetics50-C dataset. Specifically, we consider different editor ranks, ranging from 2 to 10. The results are shown in Table 5. We can observe that TTA performance improves as the rank increases. Nevertheless, thanks to the parameter efficiency, even a small rank can achieve competitiveness, relieving us from the need to exhaustively search for the optimal rank.

Impact of editing depth. We examine the influence of editing depth for representation editors on multimodal TTA performance under 5 different settings: the first layer, the last layer, the first half, the latter half, and all layers. The results are reported in Table 5. We find that, with the same number of edited layers, shallow intervention performs better than deep, which is consistent with the intuition that representations should be calibrated as early as possible. In addition, editing multiple layers outperforms editing a single layer, and all layers achieve the best performance.

Number of trainable parameters. We evaluate the number of learnable parameters of different TTA methods, including normalization layers (Tent), attention (READ), adapters (TSA), prompts (BriMPR), and our proposed editors. As shown in Table 6, since parameters in our approach mainly come from low-rank matrices, it remains parameter-efficient, even when editors are inserted in a layer-wise manner. Even with a rank as low as 2, FIRE still achieves the best performance on Kinetics50-C dataset.

Table 7: Ablation study for different components of FIRE.

Method	Kinetics50-C			VGGSound-C		
	audio	video	both	audio	video	both
Source	69.4	60.5	31.8	25.0	56.2	9.5
w/o FFT	72.8	67.5	46.1	36.7	58.0	21.6
w/o $\mathcal{L}_{cl}$	72.9	67.9	46.8	36.6	58.0	21.7
w/o $\mathcal{L}_{stat}$	71.4	63.6	36.6	33.1	56.9	15.4
w/o $\mathcal{L}_{pc}$	72.4	67.0	44.1	35.7	57.6	21.0
w/o asym.	72.5	67.2	44.5	36.5	57.8	21.5
w/ asym. (opposite)	71.4	66.9	44.0	34.8	56.7	20.8
FIRE	**73.4**	**68.6**	**47.0**	**36.9**	**58.2**	**22.0**

Ablation of FIRE. We analyze the contributions of individual components in FIRE, including the FFT, each self-training loss, and the asymmetric design for the prediction consistency loss. For the prediction consistency loss, we replace the original asymmetric weighting with either average weighting, w/o asym., or reversed asymmetric weighting, w/ asym. (opposite). The results are reported in Table 7. From Table 7, the representation editors effectively mitigate distribution shifts, and FFT along the hidden dimension further improves performance. Among the self-training objectives, the source-target statistical alignment loss $\mathcal{L}_{stat}$ yields the largest performance gain, supporting our motivation for layer-wise representation calibration. Moreover, both average weighting and the reversed asymmetric setting degrade performance, validating the effectiveness of the asymmetric strategy.

5 Conclusion

In this work, we propose FourIer Representation Editor (FIRE), a novel multimodal TTA approach that addresses distribution shifts by directly editing intermediate representations. Motivated by the observation that representation shifts emerge in shallow layers and increase with depth, FIRE performs layer-wise representation editing. Specifically, we insert an editor after each encoder transformer layer while keeping the pretrained model frozen. Subsequently, we introduce FFT to enrich the editing subspace and improve adaptation stability. Furthermore, we optimize these editors with multi-level self-training objectives to promote cross-modal semantic alignment and source-target statistical alignment. Extensive experiments demonstrate the superiority of FIRE.

Limitations: While FIRE performs representation editing across all intermediate layers, it assigns equal importance to each layer. Exploring adaptive layer selection may improve flexibility. We leave a more comprehensive investigation as future work.

AI USE STATEMENT

We used generative AI tools solely for language proofreading to improve the clarity and readability of the manuscript and did not use them for other parts. All scientific content, including ideas, methods, and experiments, etc., was generated and verified by the authors.

REFERENCES

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *CoRR*, abs/2502.17601, 2025.

Shubhankar Borse, Shreya Kadambi, Nilesh Prasad Pandey, Kartikeya Bhardwaj, Viswanath Ganapathy, Sweta Priyadarshi, Risheek Garrepalli, Rafael Esteves, Munawar Hayat, and Fatih Porikli. Foura: Fourier low-rank adaptation. In *NeurIPS*, pp. 71504–71539, 2024.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *ICLR*. OpenReview.net, 2023.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pp. 721–725. IEEE, 2020.

Mingcai Chen, Baoming Zhang, Zongbo Han, Wenyu Jiang, Yanmeng Wang, Shuai Feng, Yuntao Du, and Bingkun Bao. Test-time selective adaptation for uni-modal distribution shift in multi-modal data. In *ICML*, pp. 9711–9727. PMLR, 2025.

Lu Chi, Borui Jiang, and Yadong Mu. Fast fourier convolution. In *NeurIPS*, pp. 4479–4488, 2020.

Xianhang Chu, Xu Yang, Kun Wei, and Xi Wang. Towards robust edge model adaptation via elastic architecture search. In *AAAI*, pp. 20615–20623. AAAI Press, 2026.

James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965.

Gyanendra Das and Sai Satyam Jena. DSCA: dynamic subspace concept alignment for lifelong VLM editing. *CoRR*, abs/2604.07965, 2026.

Hao Dong, Moru Liu, Kaiyang Zhou, Eleni N. Chatzi, Juho Kannala, Cyrill Stachniss, and Olga Fink. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. *TPAMI*, 48(5):5672–5691, 2026.

Xiao Fan, Jingyan Jiang, Zhaoru Chen, Fanding Huang, Xiao Chen, Qinting Jiang, Bowen Zhang, Xing Tang, and Zhi Wang. Moetta: Test-time adaptation under mixed distribution shifts with moe-layernorm. In *AAAI*, pp. 21011–21019. AAAI Press, 2026.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. Parameter-efficient fine-tuning with discrete fourier transform. In *ICML*, pp. 14884–14901. PMLR, 2024.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R. Glass. Contrastive audio-visual masked autoencoder. In *ICLR*. OpenReview.net, 2023.

Clément Guerner, Anej Svete, Tianyu Liu, Alexander Warstadt, and Ryan Cotterell. A geometric notion of causal probing. *CoRR*, abs/2307.15054, 2023.

Zirun Guo and Tao Jin. Smoothing the shift: Towards stable test-time adaptation under complex multimodal noises. In *ICLR*. OpenReview.net, 2025.

Yongbo He, Zirun Guo, and Tao Jin. Decoupling stability and plasticity for multi-modal test-time adaptation. *CoRR*, abs/2603.00574, 2026.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*. OpenReview.net, 2019.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *ICML*, pp. 2790–2799. PMLR, 2019.

Disen Hu, Xun Jiang, Xiaofeng Cao, Zheng Wang, Jingkuan Song, Heng Tao Shen, and Xing Xu. Hyper-opinion vagueness quantification for robust multimodal learning. In *AAAI*, pp. 21831–21839. AAAI Press, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*. OpenReview.net, 2022.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *NeurIPS*, pp. 2427–2440, 2021.

Ziyu Jia, Tingyu Du, Zhengyu Tian, Hongkai Li, Yong Zhang, and Chenyu Liu. A multimodal bimamba network with test-time adaptation for emotion recognition based on physiological signals. In *NeurIPS*, pp. 145335–145358, 2025.

Qing-Yuan Jiang, Longfei Huang, and Yang Yang. Rethinking multimodal learning from the perspective of mitigating classification ability disproportion. In *NeurIPS*, pp. 129580–129607, 2025.

Will Kay, João 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. *CoRR*, abs/1705.06950, 2017.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*. OpenReview.net, 2015.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. Probing for the usage of grammatical number. In *ACL*, pp. 8818–8831. ACL, 2022.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *ICLR*. OpenReview.net, 2024.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontañón. Fnet: Mixing tokens with fourier transforms. In *NAACL*, pp. 4296–4313. ACL, 2022.

Haobin Li, Peng Hu, Qianjun Zhang, Xi Peng, Xiting Liu, and Mouxing Yang. Test-time adaptation for cross-modal retrieval with query shift. In *ICLR*. OpenReview.net, 2025.

Jiacheng Li and Songhe Feng. Bridging modalities via progressive re-alignment for multimodal test-time adaptation. In *AAAI*, pp. 22931–22939. AAAI Press, 2026.

Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *TPAMI*, 46(8):5743–5762, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, pp. 34892–34916, 2023.

Yiyang Liu, James Chenhao Liang, Ruixiang Tang, Yugyung Lee, Majid Rabbani, Sohail A. Dianat, Raghuveer Rao, Lifu Huang, Dongfang Liu, Qifan Wang, and Cheng Han. Re-imagining multimodal instruction tuning: A representation view. In *ICLR*. OpenReview.net, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, pp. 16888–16905. PMLR, 2022.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *ICLR*. OpenReview.net, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *ICML*, pp. 38298–38315. PMLR, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.

Mohammed Rakib and Arunkumar Bagavathi. G^2d: Boosting multimodal learning with gradient-guided distillation. In *ICCV*, pp. 4059–4068. IEEE, 2025.

Merey Ramazanova, Alejandro Pardo, Bernard Ghanem, and Motasem Alfarra. Test-time adaptation for combating missing modalities in egocentric videos. In *ICLR*. OpenReview.net, 2025.

V Umapathi Reddy. On fast fourier transform: a popular tool for spectrum analysis. *Resonance*, 3 (10):79–88, 1998.

Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. MM-TTA: multi-modal test-time adaptation for 3d semantic segmentation. In *CVPR*, pp. 16907–16916. IEEE, 2022.

Paul Smolensky. Neural and conceptual interpretation of pdp models. *Parallel distributed processing: Explorations in the microstructure of cognition*, 2:390–431, 1986.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models. In *ACL*, pp. 566–581. ACL, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 9(11), 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*. OpenReview.net, 2021.

Guowei Wang, Fan Lyu, and Changxing Ding. Partition-then-adapt: Combating prediction bias for reliable multi-modal test-time adaptation. In *NeurIPS*, pp. 2575–2601, 2025.

Muling Wu, Wenhao Liu, Xiaohua Wang, Tianlong Li, Changze Lv, Zixuan Ling, Jianhao Zhu, Cenyuan Zhang, Xiaoqing Zheng, and Xuanjing Huang. Advancing parameter efficiency in fine-tuning via representation editing. In *ACL*, pp. 13445–13464. ACL, 2024a.

Xiangyu Wu, Feng Yu, Yang Yang, Qing-Guo Chen, and Jianfeng Lu. Multi-label test-time adaptation with bound entropy minimization. In *ICLR*. OpenReview.net, 2025.

Xiangyu Wu, Dongming Jiang, Feng Yu, Yueying Tian, Jiaqi Tang, Qing-Guo Chen, Yang Yang, and Jianfeng Lu. Adaptive debiasing tsallis entropy for test-time adaptation. *CoRR*, abs/2602.11743, 2026.

Zhengxuan Wu, Aryaman Arora, Zheng Wang, Atticus Geiger, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. Reft: Representation finetuning for language models. In *NeurIPS*, pp. 63908–63962, 2024b.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C. H. Ngai. MENTOR: multi-level self-supervised learning for multimodal recommendation. In *AAAI*, pp. 12908–12917. AAAI Press, 2025.

Jinglin Xu, Yi Li, Chuxiong Sun, Xiao Xu, Jiangmeng Li, and Fanjiang Xu. Multi-modal test-time adaptation via adaptive probabilistic gaussian calibration. *CoRR*, abs/2604.19093, 2026.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. Test-time adaptation against multi-modal reliability bias. In *ICLR*. OpenReview.net, 2024.

Xu Yang, Moqi Li, and Kun Wei. Diffusion-calibrated continual test-time adaptation. In *AAAI*, pp. 27684–27692. AAAI Press, 2026.

Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huang, Ying Nian Wu, and Dongfang Liu. Visual fourier prompt tuning. In *NeurIPS*, pp. 5552–5585, 2024.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: test-time adaption by conservatively minimizing entropy. In *ICLR*. OpenReview.net, 2025.

Rongyu Zhang, Aosong Cheng, Yulin Luo, Gaole Dai, Huanrui Yang, Jiaming Liu, Ran Xu, Li Du, Dan Wang, and Yuan Du. Decomposing the neurons: Activation sparsity via mixture of experts for continual test time adaptation. In *AAAI*, pp. 36057–36065. AAAI Press, 2026.

Yusheng Zhao, Junyu Luo, Xiao Luo, Jinsheng Huang, Jingyang Yuan, Zhiping Xiao, and Ming Zhang. Attention bootstrapping for multi-modal test-time adaptation. In *AAAI*, pp. 22849–22857. AAAI Press, 2025.

Xiao Zheng, Chang Tang, Zhiguo Wan, Chengyu Hu, and Wei Zhang. Multi-level confidence learning for trustworthy multimodal classification. In *AAAI*, pp. 11381–11389. AAAI Press, 2023.

Andy Zou, Long Phan, Sarah Li Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *CoRR*, abs/2310.01405, 2023.

A Analysis of Representation Editing

We analyze the stability and effectiveness of the proposed method through the following two theorems.

Notation. The editing subspace $\hat{\boldsymbol{U}} \in \mathbb{R}^{d_m \times d_n}$ satisfies $\hat{\boldsymbol{U}}\hat{\boldsymbol{U}}^\top = \boldsymbol{I}_{d_m}$. Let orthogonal projection $\boldsymbol{P} = \hat{\boldsymbol{U}}^\top\hat{\boldsymbol{U}}$ and define subspace $\mathcal{S} = \mathrm{range}(\hat{\boldsymbol{U}}^\top)$. Define $\boldsymbol{h}' = \boldsymbol{W}\boldsymbol{h} + \boldsymbol{b}$.

Theorem 1 (Stability)**.** *Assuming that the frozen downstream network g is L-Lipschitz. For every $\boldsymbol{h}, \boldsymbol{h}_1, \boldsymbol{h}_2 \in \mathbb{R}^{d_n}$, then $\|\psi(\boldsymbol{h}_1) - \psi(\boldsymbol{h}_2)\|_2 \le (1 + \|\boldsymbol{W} - \hat{\boldsymbol{U}}\|_2)\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2$ and $\|g(\psi(\boldsymbol{h})) - g(\boldsymbol{h})\|_2 \le L(\|\boldsymbol{W} - \hat{\boldsymbol{U}}\|_2\|\boldsymbol{h}\|_2 + \|\boldsymbol{b}\|_2)$.*

Proof. Define $\delta = \boldsymbol{h}_1 - \boldsymbol{h}_2$. Since the bias terms cancel out in the difference, we can obtain:

$$\psi(\boldsymbol{h}_1) - \psi(\boldsymbol{h}_2) = \delta + \hat{\boldsymbol{U}}^\top(\boldsymbol{W} - \hat{\boldsymbol{U}})\delta. \tag{11}$$

By $\|\hat{\boldsymbol{U}}^\top\|_2 = 1$ and the triangle inequality, we have:

$$\|\psi(\boldsymbol{h}_1) - \psi(\boldsymbol{h}_2)\|_2 \le (1 + \|\boldsymbol{W} - \hat{\boldsymbol{U}}\|_2)\|\boldsymbol{h}_1 - \boldsymbol{h}_2\|_2. \tag{12}$$

Since the downstream network g is L-Lipschitz, we have:

$$\begin{aligned} \|g(\psi(\boldsymbol{h})) - g(\boldsymbol{h})\|_2 &\le L(\|\psi(\boldsymbol{h}) - \boldsymbol{h}\|_2) \\ &\le L(\|\boldsymbol{h}' - \hat{\boldsymbol{U}}\boldsymbol{h}\|_2) \\ &\le L(\|(\boldsymbol{W} - \hat{\boldsymbol{U}})\boldsymbol{h} + \boldsymbol{b}\|_2) \\ &\le L(\|\boldsymbol{W} - \hat{\boldsymbol{U}}\|_2\|\boldsymbol{h}\|_2 + \|\boldsymbol{b}\|_2). \end{aligned} \tag{13}$$

Near the no-op initialization, both $\boldsymbol{W} - \hat{\boldsymbol{U}}$ and $\|\boldsymbol{b}\|_2$ remain small, yielding a quantifiable upper bound on the perturbation to the frozen model.

This completes the proof. □

Theorem 2 (Wasserstein interpretation and bound)**.** *Assume that the source and target representations follow diagonal Gaussian distributions. Then, the representation statistical alignment loss and the 2-Wasserstein distance between the corresponding source and target distributions differ by at most a constant factor $\sqrt{2}$.*

Proof. To provide a probabilistic interpretation of the representation statistics alignment objective, we adopt a moment-matched diagonal Gaussian approximation for the layer-wise source and target representations. Importantly, the Gaussian assumption is used only for theoretical interpretation and is not required by the FIRE optimization procedure.

Fix a layer l and modality r. Approximate the source and target representations using the following diagonal Gaussian distributions:

$$\mathcal{N}(\mu^{\mathcal{S}}, \textstyle\sum_{\mathcal{S}})), \mathcal{N}(\mu^{\mathcal{T}}, \textstyle\sum_{\mathcal{T}}), \tag{14}$$

where $\sum_{\mathcal{S}} = \mathrm{diag}((\sigma^{\mathcal{S}})^2)$, $\sum_{\mathcal{T}} = \mathrm{diag}((\sigma^{\mathcal{T}})^2)$ and the standard deviation vectors are element-wise nonnegative. The 2-Wasserstein distance $\mathcal{W}_2$ between the Gaussian distributions is given by:

$$\begin{aligned} \mathcal{W}_2^2 &= \|\mu^{\mathcal{T}} - \mu^{\mathcal{S}}\|_2^2 + \mathrm{tr}(\textstyle\sum_{\mathcal{S}} + \sum_{\mathcal{T}} - 2(\sum_{\mathcal{T}}^{\frac{1}{2}} \sum_{\mathcal{S}} \sum_{\mathcal{T}}^{\frac{1}{2}})^{\frac{1}{2}}) \\ &= \|\mu^{\mathcal{T}} - \mu^{\mathcal{S}}\|_2^2 + \|\sigma^{\mathcal{T}} - \sigma^{\mathcal{S}}\|_2^2. \end{aligned} \tag{15}$$

Let $a = \|\mu^{\mathcal{T}} - \mu^{\mathcal{S}}\|_2$ and $b = \|\sigma^{\mathcal{T}} - \sigma^{\mathcal{S}}\|_2$. Then $\mathcal{W}_2 = \sqrt{a^2 + b^2}$ and $D_{stat} = a + b$. Since $a \ge 0$ and $b \ge 0$, we have $\sqrt{a^2 + b^2} \le a + b$. By the Cauchy–Schwarz inequality,

$$a + b \le \sqrt{2}\sqrt{a^2 + b^2}, \tag{16}$$

we can obtain:

$$\mathcal{W}_2 \le D_{stat} \le \sqrt{2}\mathcal{W}_2. \tag{17}$$

Algorithm 2 Optimization algorithm of our approach

Input: The pretrained model $F_\Theta = \{f_{\Theta^a}, f_{\Theta^v}, f_{\Theta^m}, f_{\Theta^c}\}$, representation editors $\{\psi_a^l, \psi_v^l\}_{l=1}^n$, and the unlabeled target dataset $\mathcal{D}_\mathcal{T} = \{(\boldsymbol{x}_j^a, \boldsymbol{x}_j^v)\}_{j=1}^{N_\mathcal{T}}$.
Output: The adapted editors and the prediction.

for each mini-batch in $\mathcal{D}_\mathcal{T}$ **do**
  $\hat{\boldsymbol{h}}_a^0 = \mathrm{Embed}(\boldsymbol{x}^a)$, $\hat{\boldsymbol{h}}_v^0 = \mathrm{Embed}(\boldsymbol{x}^v)$;
  for $l = 1, \ldots, n$ **do**
    // Extract representations
    $\boldsymbol{h}_a^l = f_{\Theta^a}^l(\hat{\boldsymbol{h}}_a^{l-1})$ and $\boldsymbol{h}_v^l = f_{\Theta^v}^l(\hat{\boldsymbol{h}}_v^{l-1})$;
    // Edit representations
    $\hat{\boldsymbol{h}}_a^l = \psi_a^l(\boldsymbol{h}_a^l)$ and $\hat{\boldsymbol{h}}_v^l = \psi_v^l(\boldsymbol{h}_v^l)$.
    Compute statistics $(\mu_l^{\mathcal{T},a}, \sigma_l^{\mathcal{T},a}, \mu_l^{\mathcal{T},v}, \sigma_l^{\mathcal{T},v})$;
  end for
  Obtain the prediction $\boldsymbol{p} = f_{\Theta^c}(f_{\Theta^m}(\hat{\boldsymbol{h}}_a^n, \hat{\boldsymbol{h}}_v^n))$.
  Construct modality-masked inputs and obtain predictions $\boldsymbol{p}^a$ and $\boldsymbol{p}^v$.
  Calculate the final loss in Equation 10.
  Update the editors using back-propagation.
end for

Averaging over layers and modalities preserves the direction of the inequality. Therefore, the representation statistical alignment loss and the 2-Wasserstein distance between the corresponding source and target distributions differ by at most a constant factor $\sqrt{2}$. The editors can control the optimal transport distance between the source and target domains by minimizing the loss in Equation 7.

This completes the proof. □

B Algorithm of FIRE

We summarize the overall optimization procedure of our method in Algorithm 2.

C Notation Definition

We summarize the notation definitions we used in this paper in Table I.

D Experiments Details

D.1 Dataset Details

Kinetics50: The Kinetics Kay et al. (2017) dataset is a large-scale and high-quality dataset for human action recognition in videos. The dataset consists of around 300,000 video clips covering 400 human action classes with at least 400 video clips for each action class. Each video clip lasts around 10 seconds and is labeled with a single action class. Following Yang et al. (2024), we use a subset of Kinetics which consists of 50 classes, 29,204 training pairs, and 2,466 test pairs.

VGGSound: VGGSound Chen et al. (2020) is a large-scale audio-visual correspondence dataset consisting of short clips of audio sounds, extracted from YouTube videos. All videos are captured in the wild and exhibit audio-visual correspondence, with the sound sources being visually apparent. Each video in this dataset has a fixed duration of 10 seconds.

Kinetics50-C and VGGSound-C: Following previous work Yang et al. (2024), we introduce 15 types of video corruptions and 6 types of audio corruptions. Video corruptions include "Gaussian Noise" (Gauss.), "Shot Noise" (Shot), "Impulse Noise" (Impul.), "Defocus Blur" (Defoc.), "Glass Blur" (Glass), "Motion Blur" (Motion), "Zoom Blur" (Zoom), "Snow" (Snow), "Frost" (Frost), "Fog" (Fog), "Brightness" (Bright.), "Contrast" (Contr.), "Elastic" (Elastic), "Pixelate" (Pixel) and "JPEG"

Table I: Notation Definition

Notation	Description
$N_{\mathcal{S}}$	The number of source data.
$N_{\mathcal{T}}$	The number of target data.
y_i	Category label of i-th data.
$\boldsymbol{x}_i^a/\boldsymbol{x}_i^v$	Audio / Video data point.
$\mathcal{D}_{\mathcal{S}}/\mathcal{D}_{\mathcal{T}}$	Source / Target domain.
r	r-modality.
F_Θ	Pretrained model.
$f_{\Theta^a}/f_{\Theta^v}$	Audio / Video encoder.
f_{Θ^m}	Fusion layer.
f_{Θ^c}	Classifier head.
Θ	Model parameters.
$P_{\mathcal{S}}/P_{\mathcal{T}}$	Source / Target distribution.
$\Psi(\cdot)$	Representation editor.
$\psi(\cdot)$	Fourier representation editor.
$\boldsymbol{U}/\boldsymbol{W}/\boldsymbol{b}$	Editor parameters.
$\mathcal{F}(\cdot)$	Discrete Fourier Transform.
$\mathfrak{R}$	Real component.
$\hat{\boldsymbol{U}}$	Fourier-enhanced subspace.
$\mathrm{qf}(\cdot)$	Orthogonal factor Q in QR decomposition.
B	Batch size.
$sim(\cdot,\cdot)$	Cosine similarity.
$\mathcal{L}_{cl}$	Cross-modal semantic alignment loss.
τ	Temperature hyperparameter in $\mathcal{L}_{cl}$.
μ/σ	Representation mean / standard deviation.
n	The number of encoder layers.
$\mathcal{L}_{stat}$	Source-target statistical alignment loss.
$\boldsymbol{p}$	Prediction with complete modalities.
$\boldsymbol{p}^r$	Prediction after masking modality r.
$\ell(\cdot)$	Cross-entropy loss.
$\mathcal{L}_{pc}$	Asymmetric prediction consistency loss.

(JPEG). Audio corruptions include "Gaussian Noise" (Gauss.), "Paris Traffic Noise" (Traff.), "Crowd Noise" (Crowd), "Rainy Noise" (Rain), "Thunder Noise" (Thund.) and "Windy Noise" (Wind).

D.2 Baseline Details

Tent: Tent Wang et al. (2021) optimizes the affine parameters in the normalization layer by minimizing the entropy of the model's predictions. In the transformer-based CAV-MAE, we replace the Batch Normalization (BN) layers in the original implementation with Layer Normalization (LN) layers.

T3A: T3A Iwasawa & Matsuo (2021) replaces the output layer of the predictor trained on the source domain with a pseudo-prototypical classifier and adjusts the prototype features during test time.

EATA: EATA Niu et al. (2022) selects test samples with low entropy to participate in entropy minimization, and introduces a weighted Fisher regularizer. The entropy threshold E_0 is set to $0.4 \times \ln C$ (where C is the number of classes). The cosine similarity threshold ϵ used for filtering redundant samples is set to 0.1. The trade-off hyperparameter β is set to 1. The Fisher information is calculated using 2,000 unlabeled in-distribution samples, and the moving average factor α is set to 0.1.

SAR: SAR Niu et al. (2023) introduces sharpness-aware learning and minimizes entropy. The entropy threshold E_0 is set to $0.4 \times \ln C$. The radius ρ in the sharpness-aware optimization is set to 0.05. The model recovery threshold e_0 is set to 0.2. The moving average factor used to track the loss value is set to 0.9.

DeYO: DeYO Lee et al. (2024) introduces the Pseudo-Label Probability Difference (PLPD) metric to identify harmful samples that cannot be detected by entropy. The entropy threshold τ_{Ent} is set to $0.5 \times \ln C$, the PLPD threshold τ_{PLPD} is set to $0.2 \times \ln C$, and the normalization factor Ent_0 in the weighting function is set to $0.4 \times \ln C$. The default patch-shuffling is used as the image transformation.

FOA: FOA Niu et al. (2024) inserts prompts at the input level and employs the derivative-free covariance matrix adaptation (CMA) evolution strategy. The number of prompt embeddings N_p is set to 1 with the default uniform initialization. The population size K in the CMA evolution strategy is set to $27 = 4 + 3\ln(2 \times 768)$. The trade-off parameter λ in the fitness function is set to 0.4. The step size γ in the back-to-source activation shifting is set to 1.0. The moving average factor for computing the test statistics is set to 0.1.

READ: READ Yang et al. (2024) updates the self-attention layer of the fusion module by optimizing the confidence-aware loss. The confidence threshold γ is set to e^{-1}.

ABPEM: ABPEM Zhao et al. (2025) aligns cross-attention to self-attention to reduce inter-modal differences, while excluding non-dominant class samples to reduce gradient noise in entropy loss. The threshold k for class ranking is set to 8/30 for Kinetics50-C and VGGSound-C. The weight λ for the attention bootstrapping loss is set to 1.

SuMi: SuMi Guo & Jin (2025) selects high-quality samples via unimodal-assisted identification, and balances adaptation across modalities by leveraging mutual information sharing. The multimodal threshold γ_m and the normalization factor Ent_0 are set to $0.4 \times \ln C$. The unimodal threshold γ_u is set to e^{-1}. The smoothing coefficient β is set to 0.6/0.9, the weighting term λ is set to 5.0, and the unimodal assistance t is set to 1.0 by default for Kinetics50-C and VGGSound-C. For the multimodal shift setting and real-world shift setting, we set the mutual information sharing term t_0 as $iter/2$.

TSA: TSA Chen et al. (2025) introduces modality-specific adapters and a router module to select the modality to adapt. The coefficient of the self-training loss is set to 0.5, and the softmax temperature is set to 0.001.

PTA: PTA Wang et al. (2025) consists of two key components: Partition and Debiased Reweighting (PDR) and multimodal Attention-Guided Alignment (AGA). The weight s assigned to samples with large prediction bias is set to 0.5.

BriMPR: BriMPR Li & Feng (2026) employs prompt tuning to align the global feature distribution of each modality with its corresponding source distribution. The default number of prompts per layer is set to 10. τ_0 and D_0 are set to 0.2 and 5, respectively. τ is set to 0.07/0.25 for the unimodal and multimodal corruption settings, respectively.