Bridging Modalities via Progressive Re-alignment for Multimodal Test-Time Adaptation

Jiacheng Li[1, 3], **Songhe Feng**[2, 3*]
[1]Key Laboratory of Big Data & Artificial Intelligence in Transportation (Beijing Jiaotong University), Ministry of Education, China
[2]Tangshan Research Institute, Beijing Jiaotong University, China
[3]School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
jiacheng.li@bjtu.edu.cn, shfeng@bjtu.edu.cn

Abstract

Test-time adaptation (TTA) enables online model adaptation using only unlabeled test data, aiming to bridge the gap between source and target distributions. However, in multimodal scenarios, varying degrees of distribution shift across different modalities give rise to a complex coupling effect of unimodal shallow feature shift and cross-modal high-level semantic misalignment, posing a major obstacle to extending existing TTA methods to the multimodal field. To address this challenge, we propose a novel multimodal test-time adaptation (MMTTA) framework, termed as **Bri**dging **M**odalities via **P**rogressive **R**e-alignment (**BriMPR**). BriMPR, consisting of two progressively enhanced modules, tackles the coupling effect with a divide-and-conquer strategy. Specifically, we first decompose MMTTA into multiple unimodal feature alignment sub-problems. By leveraging the strong function approximation ability of prompt tuning, we calibrate the unimodal global feature distributions to their respective source distributions, so as to achieve the initial semantic re-alignment across modalities. Subsequently, we assign the credible pseudo-labels to combinations of masked and complete modalities, and introduce inter-modal instance-wise contrastive learning to further enhance the information interaction among modalities and refine the alignment. Extensive experiments on MMTTA tasks, including both corruption-based and real-world domain shift benchmarks, demonstrate the superiority of our method.

Code — https://github.com/Luchicken/BriMPR

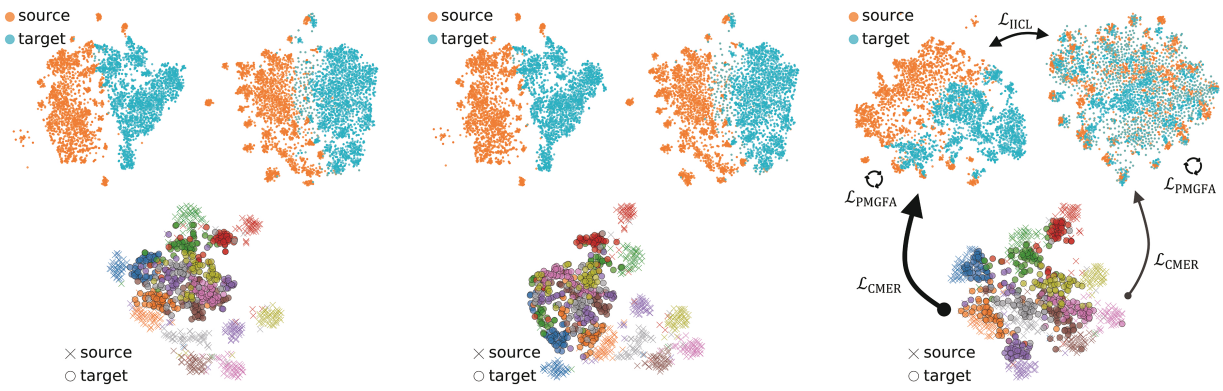
Introduction

Despite the remarkable success of deep neural networks in various fields, their excellent performances often hinge on specific data conditions. The possible distribution shift (or domain shift) between training and testing data has become a major obstacle to model generalization. Unsupervised domain adaptation (UDA) (Tzeng et al. 2014; Long et al. 2015; Jin et al. 2020) and domain generalization (DG) (Zhou et al. 2021; Li et al. 2018; Zhang and Feng 2023) have been proposed to mitigate domain gaps by designing sophisticated strategies that enable the model to adapt to the target domain during training. In contrast, test-time adaptation (TTA) (Sun et al. 2020; Wang et al. 2021; Zhang, Levine, and Finn 2022; Niu et al. 2022) adjusts the model according to specific test data during the test stage, reducing the dependence on the training process and training data, thereby making it a promising and more practical solution.

With the advancement of sensor technology, integrating and leveraging multimodal data collected from diverse sensors has significantly enhanced the perception capability of intelligent systems. Nevertheless, multimodal data also suffer from distribution shifts. What's worse, due to the complexity of multimodal data, different modalities often exhibit varying degrees of distribution shift from the source domain, inducing a complex coupling effect of unimodal shallow feature shift and cross-modal high-level semantic misalignment. Existing TTA methods, which are primarily designed for unimodal tasks, struggle to ensure consistent improvements across all modalities and often fail to fully exploit the rich information available in multimodal inputs. In Fig. 1, we visualize both unimodal and multimodal feature representations during the adaptation on the audio-visual event classification dataset Kinetics50–C (Yang et al. 2024). As a representative unimodal TTA method, EATA (Niu et al. 2022) reduces the uncertainty of model predictions by minimizing the entropy of reliable samples. However, it shows limited improvement in bridging the domain gap between source and target features for each modality. READ (Yang et al. 2024), a pioneering method for multimodal test-time adaptation (MMTTA), adapts the model by updating the self-attention layers in the fusion module to assign more weights to the high-quality modality. Nevertheless, it lacks the correction of shallow unimodal features. As shown in Fig. 1a and Fig. 1b, the lack of effective guidance for unimodal features hinders proper alignment across modalities. As a result, the fused multimodal feature representations derived from multiple unimodal features become entangled, leading to a significant decline in discriminability.

In this work, we propose **Bri**dging **M**odalities via **P**rogressive **R**e-alignment (**BriMPR**) for multimodal test-time adaptation. Through the joint efforts of self-calibration for each modality and inter-modal information interaction, BriMPR realigns the modalities that are subject to distribution shift with each other. Since the feature representations of each modality are well-aligned in the source space, we first decompose MMTTA into multiple unimodal feature

*Corresponding Author.



(a) EATA (b) READ (c) BriMPR

Figure 1: t-SNE visualizations of unimodal (top) and fused multimodal (bottom) features during adaptation versus source features. For fused features, 10 classes from Kinetics50–C are shown.

alignment sub-problems. Leveraging the strong function approximation ability of prompt tuning (Wang et al. 2023), we calibrate the global feature distribution of each modality to its corresponding source distribution via modality-specific prompts embedded across layers of the modality-specific encoders, thereby indirectly achieving initial cross-modal semantic alignment. Subsequently, the alignment is further refined by enhancing inter-modal information interaction. We propose a novel cross-modal masked embedding recombination loss, which promotes the extraction of multimodal information by providing calibrated pseudo-labels for the combinations of masked and complete modalities. Additionally, we introduce inter-modal instance-wise contrastive learning to maintain cross-modal alignment at the instance level. As shown in Fig. 1c, BriMPR effectively bridges the domain gap between the source and target for each unimodal feature, thereby enhancing the discriminability of the fused features. Our contributions can be summarized as follows:

- We propose a novel MMTTA framework which mitigates modality-wise distribution shifts in a divide-and-conquer manner, facilitating the re-alignment among modalities.
- We leverage the excellent function approximation ability of prompt tuning to achieve efficient calibration of the unimodal global feature distribution, and propose a novel cross-modal masked embedding recombination strategy to enhance the inter-modal interaction.
- We conduct extensive experiments on MMTTA benchmarks, including corruption shift and real-world shift datasets, demonstrating the superiority of BriMPR over existing SOTA methods.

Related Work

Test-Time Adaptation. Test-time adaptation (TTA) leverages unlabeled test data to adapt models to unseen target domains during test-time. The idea of TTA can be traced back to TTT (Sun et al. 2020), which uses a self-supervised auxiliary branch to enable adaptation during inference. A series of works (Wang et al. 2021; Niu et al. 2022, 2023; Lee et al. 2024) explore fully test-time adaptation (FTTA) by optimizing the normalization layers via entropy-based losses, without altering the pre-training stage. Given the limitations of unimodal TTA methods in multimodal scenarios, MM-TTA (Shin et al. 2022) proposes a cross-modal self-learning framework for MMTTA. READ (Yang et al. 2024) highlights the reliability bias of MMTTA under unimodal corruption, and proposes to adaptively assign modality weights by optimizing the self-attention in the fusion module. ABPEM (Zhao et al. 2025) reduces the gap between cross-attention and self-attention, and computes the principal part of entropy to reduce gradient noise. SuMi (Guo and Jin 2025) utilizes interquartile range smoothing to identify samples used for calculating entropy loss. Moreover, AEO (Dong, Chatzi, and Fink 2025) introduces unseen classes and proposes the Multimodal open-set test-time adaptation setting. In this work, we attribute the difficulties of MMTTA to the coupling effect of unimodal shallow feature shift and cross-modal high-level semantic misalignment, and propose a divide-and-conquer method to re-bridge modalities during testing.

Prompt Tuning. Originally developed in natural language processing, prompt tuning introduces extra tokens to guide models toward generating task-specific outputs. In computer vision, approaches like CoOp (Zhou et al. 2022b) and CoCoOp (Zhou et al. 2022a) leverage learnable prompts to enhance the zero-shot recognition capabilities of vision-language models (VLMs). Integrating the idea of TTA, test-time prompt tuning (TPT) (Shu et al. 2022; Feng et al. 2023; Zhang et al. 2024a) fine-tunes text prompts using test samples to improve the generalization of VLMs. While TPT primarily focuses on extracting rich knowledge from large-scale VLMs, our work is more closely aligned with visual prompt tuning (VPT) (Jia et al. 2022; Yoo et al. 2023). VPT introduces prompt tuning into Vision Transformer, achieving significant performance gains over full fine-tuning. Our work extends prompt tuning to MMTTA tasks, leveraging the strong function approximation ability of prompts to efficiently calibrate the distribution of each unimodal feature—not limited to visual features alone.

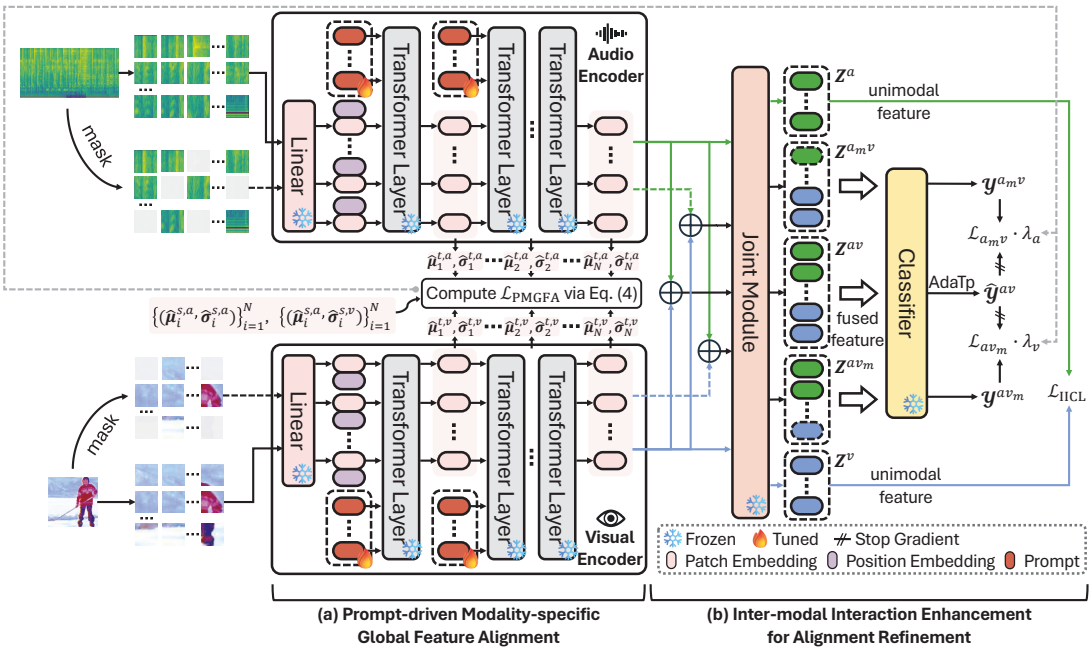


Figure 2: Overview of BriMPR. BriMPR achieves initial alignment and alignment refinement through two progressive modules. The added modality-specific prompts are used to project the unimodal features into the re-aligned feature space.

Preliminaries

Multimodal Test-Time Adaptation (MMTTA). Without loss of generality, we take two modalities as an example to provide a formal definition of MMTTA. An off-the-shelf model $\mathcal{F}_\Theta$ pre-trained on the source domain $\mathcal{D}_\mathcal{S} = \{(\mathbf{x}_i^{u_1}, \mathbf{x}_i^{u_2}, y_i)\}_{i=1}^{N_\mathcal{S}}$ is adopted as the initial model, where the two modalities of the source data follow the probability distributions $\mathbf{x}_i^{u_1} \sim P_{\mathcal{S},u_1}(\mathbf{x})$ and $\mathbf{x}_i^{u_2} \sim P_{\mathcal{S},u_2}(\mathbf{x})$, respectively. The goal of MMTTA is to adapt $\mathcal{F}_\Theta$ online to the target domain $\mathcal{D}_\mathcal{T} = \{(\mathbf{x}_j^{u_1}, \mathbf{x}_j^{u_2})\}_{j=1}^{N_\mathcal{T}}$, where the two modalities of target data follow the probability distributions $\mathbf{x}_j^{u_1} \sim P_{\mathcal{T},u_1}(\mathbf{x})$ and $\mathbf{x}_j^{u_2} \sim P_{\mathcal{T},u_2}(\mathbf{x})$. During adaptation, the source domain is inaccessible and there is a domain shift between the source and target distributions, i.e., $P_{\mathcal{S},u_1}(\mathbf{x}) \neq P_{\mathcal{T},u_1}(\mathbf{x})$ and $P_{\mathcal{S},u_2}(\mathbf{x}) \neq P_{\mathcal{T},u_2}(\mathbf{x})$.

Prompt Tuning. Prompt tuning is regarded as a parameter-efficient fine-tuning technique, which adapts the model to downstream tasks by prepending and optimizing learnable prompt tokens into the input sequence (Li and Liang 2021; Lester, Al-Rfou, and Constant 2021; Jia et al. 2022; Liu et al. 2022). For an encoder Φ consisting of N transformer layers, when inserting a specified number of prompts into the input sequence at each layer, the forward process of the i-th layer can be formulated as:

$$[\_; \boldsymbol{E}_i] = L_i\left([\boldsymbol{P}_{i-1}; \boldsymbol{E}_{i-1}]\right), \quad i = 1, ..., N. \tag{1}$$

Here $\boldsymbol{E}_i = [\boldsymbol{e}_{i,1}; \boldsymbol{e}_{i,2}; \ldots; \boldsymbol{e}_{i,m}]$ and $\boldsymbol{P}_i = [\boldsymbol{p}_{i,1}; \boldsymbol{p}_{i,2}; \ldots; \boldsymbol{p}_{i,m_p}]$ denote the sequences of original input tokens and inserted prompt tokens, where m and m_p is the number of tokens, and the token dimension is d. $[\cdot;\cdot]$ denotes token-level concatenation. Then, a supervised loss $\mathcal{L}$ is minimized over the downstream dataset $\mathcal{D}_{\text{ds}}$ to obtain the optimal prompt $\boldsymbol{P}^* = \{\boldsymbol{P}_0^*, \boldsymbol{P}_1^*, \ldots, \boldsymbol{P}_{N-1}^*\}$:

$$\boldsymbol{P}^* = \arg\min_{\boldsymbol{P}} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_{\text{ds}}} \mathcal{L}(h(\text{MeanPool}(\boldsymbol{E}_N)), y), \tag{2}$$

where h denotes the classifier. In MMTTA, due to the absence of annotation for the test data, the loss must be reformulated to enable the learning of task-specific prompts.

Methodology

In this section, we introduce BriMPR for MMTTA, with its overall framework illustrated in Fig. 2. BriMPR comprises two progressively enhanced modules: (a) *Prompt-driven Modality-specific Global Feature Alignment* achieves initial cross-modal alignment by minimizing the discrepancy between the unimodal target statistics and their corresponding in-distribution statistics; (b) *Inter-modal Interaction Enhancement for Alignment Refinement* further refines the alignment by providing credible pseudo-labels for combinations of masked and complete modalities, and conducting inter-modal instance-wise contrastive learning.

Following READ (Yang et al. 2024), we decompose the source model into two modality-specific encoders (Φ^a for the audio modality and Φ^v for the visual modality), a joint module Ψ, and a classifier h. We update only the prompts for each modality-specific encoder, keeping the rest of the model frozen, to recalibrate individual feature distributions and achieve bottom-up modality re-alignment.

Prompt-driven Modality-specific Global Feature Alignment (PMGFA)

The final prediction of a multimodal model comes from the joint effect of multiple individual modalities. This naturally

allows MMTTA to be decomposed into multiple unimodal test-time adaptation problems. On the other hand, if the target representations at test time can be well projected back to the corresponding source representations, then a TTA model tends to perform well. Based on the intuitions above, we decouple MMTTA into multiple modality-specific feature alignment sub-problems. Since the inter-modal semantic representations are well aligned in the source representation space, solving these sub-problems means indirectly achieving cross-modal semantic alignment of the target representation.

Concretely, we first model the modality-specific source and target feature distributions as multivariate Gaussian distributions, i.e., $P_{\mathcal{S},u} = \mathcal{N}(\mu^{s,u}, \Sigma^{s,u})$ and $P_{\mathcal{T},u} = \mathcal{N}(\mu^{t,u}, \Sigma^{t,u})$, where $u \in \{a, v\}$. In prior works (Liu et al. 2021; Su, Xu, and Jia 2022; Zhang et al. 2024b), feature alignment is typically achieved by matching the first and second moments between distributions (i.e., $||\mu^t - \mu^s||_2^2 + ||\Sigma^t - \Sigma^s||_F^2$) or minimizing the KL-divergence (i.e., $D_{KL}(P_{\mathcal{S}}||P_{\mathcal{T}})$). However, both approaches rely on the estimation of the covariance matrix Σ, whose error is significantly amplified in high-dimensional data. Therefore, we propose to retain only the diagonal elements of Σ, which reduces the estimation error by a factor of d, as supported by the following theorem:

Theorem 1. *Given $x_1, \ldots, x_n \in \mathbb{R}^d$ independently drawn from a multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$, let $\hat{\Sigma}$ be the unbiased sample covariance matrix and $\hat{\sigma}^2 = [\hat{\sigma}_1^2, \ldots, \hat{\sigma}_d^2]^T$ be the vector of its diagonal entries. Then, the mean squared errors satisfy:*

$$\mathbb{E}\left[\|\hat{\Sigma} - \Sigma\|_F^2\right] = \mathcal{O}\left(\frac{d^2}{n}\right), \mathbb{E}\left[\|\hat{\sigma}^2 - \sigma^2\|_2^2\right] = \mathcal{O}\left(\frac{d}{n}\right). \tag{3}$$

Due to space limitations, the corresponding proof can be found in Appendix. Emerging research (Wang et al. 2023) has shown that prompt tuning can serve as universal approximators for sequence-to-sequence functions. Motivated by this, we employ prompts as an implicit mapping from the target feature space to the source feature space. For the data $\mathbf{x}^u$ and the i-th layer of the modality-specific encoder Φ^u, the input sequence $\boldsymbol{E}_{i-1}^u(\mathbf{x}^u)$ undergoes attention interaction with the added prompts $\boldsymbol{P}_{i-1}^u$ to obtain the transformed output sequence $\boldsymbol{E}_i^u(\mathbf{x}^u)$. The global feature representation can be expressed as $\boldsymbol{Z}_i^u(\mathbf{x}^u) = \mathrm{MeanPool}(\boldsymbol{E}_i^u(\mathbf{x}^u))$. Subsequently, we minimize the following empirical risk on the current batch $\{(\mathbf{x}_j^a, \mathbf{x}_j^v)\}_{j=1}^B$:

$$\begin{aligned}\mathcal{L}_{\mathrm{PMGFA}} &= \sum_{u\in\{a,v\}} \mathrm{Disc}\left(P_{\mathcal{S},u}, P_{\mathcal{T},u}\right) \\ &= \sum_{u\in\{a,v\}} \frac{1}{N}\sum_{i=1}^{N}\left(\left\|\hat{\boldsymbol{\mu}}_i^{t,u} - \hat{\boldsymbol{\mu}}_i^{s,u}\right\|_2 + \left\|\hat{\boldsymbol{\sigma}}_i^{t,u} - \hat{\boldsymbol{\sigma}}_i^{s,u}\right\|_2\right),\end{aligned} \tag{4}$$

where $\mathrm{Disc}(\cdot,\cdot)$ denotes the mean of the layer-wise distribution discrepancy. For convenience, we will interchangeably use Disc^u and $\mathrm{Disc}(P_{\mathcal{S},u}, P_{\mathcal{T},u})$ in the following context. $\|\cdot\|_2$ denotes the Euclidean norm. $\hat{\boldsymbol{\mu}}_i^{t,u} = \sum_{j=1}^B \boldsymbol{Z}_i^u(\mathbf{x}_j^u)/B$ and $\hat{\boldsymbol{\sigma}}_i^{t,u} = \sqrt{\sum_{j=1}^B [(\boldsymbol{Z}_i^u(\mathbf{x}_j^u) - \hat{\boldsymbol{\mu}}_i^{t,u})^2]/(B-1)}$ are the estimated mean and standard deviation, respectively. Similar to many other TTA methods (Niu et al. 2022; Döbler, Marsden, and Yang 2023; Wang et al. 2025), we pre-compute $\{\hat{\boldsymbol{\mu}}_i^{s,u}, \hat{\boldsymbol{\sigma}}_i^{s,u}\}_{i=1}^N$ offline prior to the test phase, and this process is performed only once.

Inter-modal Interaction Enhancement for Alignment Refinement

After initial cross-modal semantic alignment via unimodal feature calibration, we further improve the quality of alignment by inter-modal interactions. By recombining masked and complete modalities, the unmasked low-quality modality is forced to draw multimodal information from credible pseudo-labels. Meanwhile, inter-modal instance-wise contrastive learning is applied to strengthen the alignment across instances.

Cross-modal Masked Embedding Recombination. Masked language modeling (Devlin et al. 2019) and masked image modeling (He et al. 2022) force model to reconstruct the masked regions by utilizing contextual clues and have been widely used as powerful self-supervised learning paradigms in natural language processing and computer vision tasks, respectively. Related but distinct, our proposed Cross-modal Masked Embedding Recombination (CMER) uses masking to simulate distribution shifts from missing patches, serving as a form of data augmentation.

For input $\mathbf{x}^u$, we randomly mask a portion (e.g., 50%) of its patches and encode the unmasked part $\mathbf{x}^{u_m}$ using Φ^u with modality-specific prompts $\boldsymbol{P}^u$ to obtain the masked embedding $\Phi^u(\mathbf{x}^{u_m})$. Then, $\Phi^u(\mathbf{x}^{u_m})$ is recombined with complete embeddings from other modalities and passed to the joint module, generating an augmented representation that simulates unimodal corruption. Taking the masked audio modality as an example, the recombined representations and their predictions are formulated as:

$$\begin{aligned}\boldsymbol{Z}^{a_m v} &= \Psi([\Phi^a(\mathbf{x}^{a_m}); \Phi^v(\mathbf{x}^v)]), \\ \boldsymbol{y}^{a_m v} &= \sigma(h(\mathrm{MealPool}(\boldsymbol{Z}^{a_m v}))),\end{aligned} \tag{5}$$

where σ denotes the softmax function. With the initial alignment from PMGFA, we can utilize the complete multimodal data to provide reliable pseudo-labels for augmented inputs. As pseudo-labels become more reliable in the later stages of adaptation, we further calibrate them via temperature scaling (Hinton, Vinyals, and Dean 2015; Guo et al. 2017):

$$\hat{\boldsymbol{y}}_k^{av} = \frac{\exp([h(\mathrm{MealPool}(\boldsymbol{Z}^{av}))]_k / \mathrm{AdaTp})}{\sum_{k'=1}^C \exp([h(\mathrm{MealPool}(\boldsymbol{Z}^{av}))]_{k'} / \mathrm{AdaTp})}. \tag{6}$$

Here, k and k' denote the k-th and k'-th elements of the tensor, and C represents the number of classes. $\mathrm{AdaTp} = 1 + \tau_0/(1 + \exp(D_0 - \mathrm{Disc}^J)) \in (1, 1+\tau_0)$ is the adaptive temperature coefficient, where Disc^J is the distribution discrepancy calculated for the joint module, and τ_0 and D_0 are predefined hyperparameters. When Disc^J is large, AdaTp approaches $1 + \tau_0$ to alleviate overconfident predictions. As Disc^J decreases, AdaTp approaches 1, and Eq. (6) approximates the vanilla softmax function. Subsequently, minimize

Method	Noise			Blur				Weather				Digital				
	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
• Tent$_{ICLR2021}$	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
• EATA$_{ICML2022}$	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
• SAR$_{ICLR2023}$	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
• DeYO$_{ICLR2024}$	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
• FOA$_{ICML2024}$	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
• READ$^{\dagger}_{ICLR2024}$	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
• ABPEM$^{\dagger}_{AAAI2025}$	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
• SuMi$^{\dagger}_{ICLR2025}$	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
• BriMPR†	**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**
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
• Tent$_{ICLR2021}$	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
• EATA$_{ICML2022}$	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
• SAR$_{ICLR2023}$	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
• DeYO$_{ICLR2024}$	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
• FOA$_{ICML2024}$	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
• READ$^{\dagger}_{ICLR2024}$	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
• ABPEM$^{\dagger}_{AAAI2025}$	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
• SuMi$^{\dagger}_{ICLR2025}$	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
• BriMPR†	**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**

Table 1: Comparison with SOTA methods on Kinetics50-C (top) and VGGSound-C (bottom) under the unimodal shift setting (severity level 5 of video corruption). †Multimodal test-time adaptation methods.

the cross-entropy between the calibrated pseudo-label and the augmented predictions:

$$\begin{aligned}\mathcal{L}_{\text{CMER}} &= \lambda^a \mathcal{L}_{a_m v} + \lambda^v \mathcal{L}_{a v_m} \\ &= -\lambda^a \sum_{k=1}^{C} \hat{\boldsymbol{y}}_k^{av} \log \boldsymbol{y}_k^{a_m v} - \lambda^v \sum_{k=1}^{C} \hat{\boldsymbol{y}}_k^{av} \log \boldsymbol{y}_k^{a v_m},\end{aligned} \tag{7}$$

where $\lambda^u = 1 - \text{Disc}^u/(\text{Disc}^a + \text{Disc}^v)$ $(u \in \{a, v\})$ is the weight of the corresponding term, assigning a higher weight to the augmentation with a milder distribution shift in the masked modality. Intuitively, $\mathcal{L}_{\text{CMER}}$ deliberately discards high-quality modality information, forcing the corrupted modality to independently derive the correct result.

Inter-modal Instance-wise Contrastive Learning. Contrastive learning (He et al. 2020; Chen et al. 2020b) has emerged as a key paradigm in cross-modal representation learning, aiming to improve the quality of representations by aligning the feature spaces of the same semantic instance across different modalities/views. Building upon the calibration of unimodal feature distributions, BriMPR introduces inter-modal instance-wise contrastive learning. For data $\mathbf{x}^u$ $(u \in \{a, v\})$, its unimodal representation is as follows:

$$\boldsymbol{Z}^u = \Psi(\Phi^u(\mathbf{x}^u)). \tag{8}$$

Subsequently, different unimodal representations of the same instance are regarded as positive pairs, while the others as negative pairs. The contrastive loss is defined as:

$$\mathcal{L}_{\text{IICL}} = -\frac{1}{2B} \sum_{j=1}^{B} \sum_{u_1 \neq u_2} \log \frac{e^{\text{sim}(\boldsymbol{Z}_j^{u_1}, \boldsymbol{Z}_j^{u_2})/\tau}}{\sum_{j'=1}^{B} e^{\text{sim}(\boldsymbol{Z}_j^{u_1}, \boldsymbol{Z}_{j'}^{u_2})/\tau}}, \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ denotes the cosine similarity function, and τ denotes the temperature hyperparameter.

Overall Procedure

To brief, BriMPR optimizes the added modality-specific prompts by minimizing the following loss:

$$\mathcal{L}_{\text{BriMPR}} = \mathcal{L}_{\text{PMGFA}} + \mathcal{L}_{\text{CMER}} + \mathcal{L}_{\text{IICL}}. \tag{10}$$

Experiments

Experimental Setups

Datasets and models. We evaluate our method on four commonly used multimodal datasets, including Kinetics50-C, VGGSound-C (Yang et al. 2024), CMU-MOSI (Zadeh et al. 2016), and CH-SIMS (Yu et al. 2020). Kinetics50-C/VGGSound-C contain two modalities: video and audio, and are obtained by adding various corruptions to the test sets of the original versions (i.e., Kinetics (Kay et al. 2017) and VGGSound (Chen et al. 2020a)). For the video modality and the audio modality, 15 and 6 types of corruption are introduced, respectively, which are divided into 5 severity levels. Following (Yang et al. 2024), we use the pre-trained CAV-MAE (Gong et al. 2023) as the source model. CMU-MOSI/CH-SIMS contain three modalities: text, video, and audio. Following (Guo and Jin 2025), we use stacked Transformer blocks as the backbone and pre-train the model on MOSI and SIMS, respectively.

Considered settings. For domain shifts caused by corruptions, we consider two tasks and report average classification accuracy (%): (1) Under the unimodal shift setting, following (Yang et al. 2024), one modality is corrupted while the other modality remains clean; (2) Under the multimodal shift setting, both modalities are corrupted. For real-world

Method	Noise			Weather			Avg.	Noise			Weather			Avg.
	Gauss.	Traff.	Crowd	Rain	Thund.	Wind		Gauss.	Traff.	Crowd	Rain	Thund.	Wind	
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
• Tent$_{ICLR2021}$	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
• EATA$_{ICML2022}$	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
• SAR$_{ICLR2023}$	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
• DeYO$_{ICLR2024}$	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
• FOA$_{ICML2024}$	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
• READ$^{\dagger}{}_{ICLR2024}$	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
• ABPEM$^{\dagger}{}_{AAAI2025}$	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
• SuMi$^{\dagger}{}_{ICLR2025}$	**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
• BriMPR†	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**

Table 2: Comparison with SOTA methods on Kinetics50-C (left) and VGGSound-C (right) under the unimodal shift setting (severity level 5 of audio corruption).

Method	Noise			Blur				Weather				Digital				
	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
• Tent$_{ICLR2021}$	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
• EATA$_{ICML2022}$	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
• SAR$_{ICLR2023}$	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
• DeYO$_{ICLR2024}$	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
• FOA$_{ICML2024}$	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
• READ$^{\dagger}{}_{ICLR2024}$	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
• ABPEM$^{\dagger}{}_{AAAI2025}$	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
• SuMi$^{\dagger}{}_{ICLR2025}$	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
• BriMPR†	**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**

Table 3: Comparison with SOTA methods on Kinetics50-C under the multimodal shift setting (severity level 5).

Method	Noise			Weather			Avg.
	Gauss.	Traff.	Crowd	Rain	Thund.	Wind	
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
• 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
• BriMPR†	**23.5**	**18.8**	**21.4**	**15.8**	**26.8**	**18.3**	**20.7**

Table 4: Comparison with SOTA methods on VGGSound-C under the multimodal shift setting (severity level 5).

domain shifts, we consider the settings of MOSI → SIMS and SIMS → MOSI, and report accuracy (ACC) and F1 score (F1).

Baselines. We compare the proposed method with multiple baselines including Source (source pre-trained model), Tent (Wang et al. 2021), EATA (Niu et al. 2022), SAR (Niu et al. 2023), DeYO (Lee et al. 2024), FOA (Niu et al. 2024), READ (Yang et al. 2024), ABPEM (Zhao et al. 2025) and SuMi (Guo and Jin 2025).

Method	MOSI → SIMS		SIMS → MOSI	
	ACC↑	F1↑	ACC↑	F1↑
Source	46.0	45.6	59.0	73.6
• Tent	38.1	42.2	59.6	74.5
• READ†	32.4	44.5	59.7	74.7
• SuMi†	44.4	45.0	59.4	74.2
• BriMPR†	**58.2**	**57.6**	**59.9**	**74.9**

Table 5: Comparison with SOTA methods on real-world shift datasets.

Implementation details. For all experiments, we use an Adam optimizer with a learning rate of 1e-4 and batch size of 64. The default number of prompts per layer m_p is set to 10 and the prompts are randomly initialized (Jia et al. 2022). The mask ratio is set to 0.5. τ_0 and D_0 of the adaptive temperature coefficient AdaTp are set to 0.2 and 5 respectively. τ in Eq. (9) is set to 0.07/0.25 for the unimodal and multimodal corruption settings respectively. For the hyperparameters of the compared methods, we adopt the recommended values from the respective papers. All the experiments are conducted with 3 random seeds on RTX-3090 GPUs.

Method	Kinetics50-C			VGGSound-C		
	audio	video	both	audio	video	both
BriMPR w/o $\mathcal{L}_{\text{CMER}}$	71.4	65.6	40.7	35.3	57.6	20.2
• BriMPR ($\lambda^a \leftrightarrow \lambda^v$)	70.0 (-1.4)	65.2 (-0.4)	39.9 (-0.8)	32.1 (-3.2)	56.5 (-1.1)	19.5 (-0.7)
• BriMPR	72.0 (+0.6)	65.9 (+0.3)	40.9 (+0.2)	36.5 (+1.2)	57.7 (+0.1)	20.7 (+0.5)

Table 6: Verify the effect of CMER from the perspective of weights.

Method	Kinetics50-C			VGGSound-C		
	audio	video	both	audio	video	both
Source	69.4	60.5	31.8	25.0	56.2	9.5
$\mathcal{L}_{\text{KL}}$	69.3	60.4	31.5	24.8	55.7	9.1
$\mathcal{L}_{\text{moment}_2}$	69.9	61.5	34.5	25.2	48.9	12.1
$\mathcal{L}_{\text{moment}_1}$	71.3	63.5	37.4	32.0	54.7	16.4
(**A**) $\mathcal{L}_{\text{PMGFA}}$	71.1	64.7	40.5	35.1	57.5	20.1
(**B**) + $\mathcal{L}_{\text{IICL}}$	71.4	65.6	40.7	35.3	57.6	20.2
(**C**) + $\mathcal{L}_{\text{CMER}}$	**72.0**	**65.9**	**40.9**	**36.5**	**57.7**	**20.7**

Table 7: Ablation studies for different components of BriMPR. $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_{\text{moment}_2}$ and $\mathcal{L}_{\text{moment}_1}$ respectively denote replacing $\mathcal{L}_{\text{PMFGA}}$ with the KL-divergence, moment matching, and moment matching in a non-squared form.

Performance Comparison

Results of the unimodal shift setting. In Tab. 1 and Tab. 2, we present the results of the unimodal shift setting on Kinetics50-C and VGGSound-C with audio corruption and video corruption, respectively. Our proposed method BriMPR consistently improves the source model and outperforms all other competing methods. Notably, in scenarios where the dominant modality of the dataset is corrupted (for Kinetics50-C, video is the dominant modality; for VGGSound-C, audio is the dominant modality), BriMPR yields significant performance gains (60.5% $\rightarrow$ 65.9% on Kinetics50-C; 25.0% $\rightarrow$ 36.5% on VGGSound-C).

Results of the multimodal shift setting. Tab. 3 and Tab. 4 respectively present the results of the challenging multimodal shift setting on Kinetics50-C and VGGSound-C. Taking the "Gauss." column in Tab. 3 as an example, the reported value denotes the average classification accuracy (%) across all 6 types of audio corruption, given the presence of Gaussian corruption in the video modality. Most methods suffer significant performance drops under this setting, whereas our BriMPR achieves the best results on most domains by decoupling MMTTA into unimodal alignment subproblems, thereby reducing the dependence on high-quality modalities.

Results of the real-world shift setting. Tab. 5 presents the results from the MOSI/SIMS datasets using text, video, and audio modalities. BriMPR exhibits strong robustness to real-world shifts. Notably, only BriMPR achieves results better than random guess ($> 50\%$) on the MOSI $\rightarrow$ SIMS task, thanks to its modulation of the target feature space.

Ablation Studies

Scrutinize CMER from the perspective of the weight λ^u. To illustrate how multimodal test-time adaptation benefits from CMER, we swap the weights λ^u ($u \in \{a, v\}$) in $\mathcal{L}_{\text{CMER}}$, assigning lower weight to the augmentation with a milder distribution shift in the masked modality. As reported in Tab. 6, the mismatched weights lead to significant performance drops. Taking the case of audio corruption as an example (where $\lambda^v > \lambda^a$), the performance degradation can be attributed to two main factors: (1) For $\lambda^a \mathcal{L}_{av_m}$, the small λ^a suppresses the extraction of multimodal information by the complete but low-quality audio modality; (2) For $\lambda^v \mathcal{L}_{a_m v}$, providing pseudo-labels to the augmentation with the masked audio modality introduces more error information into the unmasked high-quality video modality.

Component analysis. As shown in Tab. 7, we conducted an ablation study on the components of BriMPR. First, we verify the effectiveness of $\mathcal{L}_{\text{PMGFA}}$ (**A**): compared with KL-divergence (Row 2) and moment matching (Row 3), $\mathcal{L}_{\text{PMGFA}}$ demonstrates a significant advantage, as it eliminates the off-diagonal elements in the covariance matrix, reducing the estimation error. When moment matching is modified to a non-squared form (Row 4), performance improves in most cases, as the squared norm also amplifies the error. Subsequently, combining $\mathcal{L}_{\text{PMGFA}}$ (**A**), which serves as the initial alignment objective, with inter-modal instance-wise contrastive learning $\mathcal{L}_{\text{IICL}}$ (**B**) and cross-modal masked embedding recombination $\mathcal{L}_{\text{CMER}}$ (**C**) for alignment refinement, leads to further performance gains across all tasks.

Conclusion

In this paper, we introduce BriMPR, a novel MMTTA method which tackles the coupling effect of unimodal feature shift and cross-modal semantic misalignment in a divide-and-conquer manner. Specifically, benefiting from the well-aligned source feature space, we first calibrate each unimodal global feature distribution via modality-specific prompts to achieve initial cross-modal semantic alignment. We then introduce a novel Cross-modal Masked Embedding Recombination strategy to facilitate the integration of multimodal information into low-quality modalities, and further refine the alignment via Inter-modal Instance-wise Contrastive Learning. Extensive experiments conducted on MMTTA benchmark, which includes corruption datasets and real-world shift datasets, demonstrate the superiority of BriMPR over the SOTA methods.

Acknowledgments

This work was supported by the Fundamental Research Funds for the Central Universities (No. 2025JBZX059), the Natural Science Foundation of Hebei Province (No. F2025105018), the Tangshan Municipal Science and Technology Plan Project (No.23130225E) and the Beijing Natural Science Foundation (No.4242046).

References

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020a. Vggsound: A Large-Scale Audio-Visual Dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 721–725.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020b. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1597–1607. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Döbler, M.; Marsden, R. A.; and Yang, B. 2023. Robust Mean Teacher for Continual and Gradual Test-Time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7704–7714.

Dong, H.; Chatzi, E.; and Fink, O. 2025. Towards Robust Multimodal Open-set Test-time Adaptation via Adaptive Entropy-aware Optimization. In *International Conference on Learning Representations*.

Feng, C.-M.; Yu, K.; Liu, Y.; Khan, S.; and Zuo, W. 2023. Diverse Data Augmentation with Diffusions for Effective Test-time Prompt Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2704–2714.

Gong, Y.; Rouditchenko, A.; Liu, A. H.; Harwath, D.; Karlinsky, L.; Kuehne, H.; and Glass, J. R. 2023. Contrastive Audio-Visual Masked Autoencoder. In *International Conference on Learning Representations*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1321–1330. PMLR.

Guo, Z.; and Jin, T. 2025. Smoothing the Shift: Towards Stable Test-Time Adaptation under Complex Multimodal Noises. In *International Conference on Learning Representations*.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. arXiv:1503.02531.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual Prompt Tuning. In *European Conference on Computer Vision*, 709–727. Cham: Springer Nature Switzerland.

Jin, Y.; Wang, X.; Long, M.; and Wang, J. 2020. Minimum Class Confusion for Versatile Domain Adaptation. In *European Conference on Computer Vision*, 464–480. Cham: Springer International Publishing.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; Suleyman, M.; and Zisserman, A. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950.

Lee, J.; Jung, D.; Lee, S.; Park, J.; Shin, J.; Hwang, U.; and Yoon, S. 2024. Entropy is not Enough for Test-Time Adaptation: From the Perspective of Disentangled Factors. In *International Conference on Learning Representations*.

Lester, B.; Al-Rfou, R.; and Constant, N. 2021. The Power of Scale for Parameter-Efficient Prompt Tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 3045–3059. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Li, D.; Yang, Y.; Song, Y.-Z.; and Hospedales, T. 2018. Learning to Generalize: Meta-Learning for Domain Generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Li, X. L.; and Liang, P. 2021. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4582–4597. Online: Association for Computational Linguistics.

Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; and Tang, J. 2022. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 61–68. Dublin, Ireland: Association for Computational Linguistics.

Liu, Y.; Kothari, P.; van Delft, B.; Bellot-Gurlet, B.; Mordan, T.; and Alahi, A. 2021. TTT++: When Does Self-Supervised Test-Time Training Fail or Thrive? In *Advances in Neural Information Processing Systems*, volume 34, 21808–21820. Curran Associates, Inc.

Long, M.; Cao, Y.; Wang, J.; and Jordan, M. 2015. Learning Transferable Features with Deep Adaptation Networks. In *International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, 97–105. Lille, France: PMLR.

Niu, S.; Miao, C.; Chen, G.; Wu, P.; and Zhao, P. 2024. Test-Time Model Adaptation with Only Forward Passes. In *International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 38298–38315. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient Test-Time Model Adaptation without Forgetting. In *International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 16888–16905. PMLR.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-time Adaptation in Dynamic Wild World. In *International Conference on Learning Representations*.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. MM-TTA: Multi-Modal Test-Time Adaptation for 3D Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16928–16937.

Shu, M.; Nie, W.; Huang, D.-A.; Yu, Z.; Goldstein, T.; Anandkumar, A.; and Xiao, C. 2022. Test-Time Prompt Tuning for Zero-Shot Generalization in Vision-Language Models. In *Advances in Neural Information Processing Systems*, volume 35, 14274–14289. Curran Associates, Inc.

Su, Y.; Xu, X.; and Jia, K. 2022. Revisiting Realistic Test-Time Training: Sequential Inference and Adaptation by Anchored Clustering. In *Advances in Neural Information Processing Systems*, volume 35, 17543–17555. Curran Associates, Inc.

Sun, Y.; Wang, X.; Liu, Z.; Miller, J.; Efros, A.; and Hardt, M. 2020. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 9229–9248. PMLR.

Tzeng, E.; Hoffman, J.; Zhang, N.; Saenko, K.; and Darrell, T. 2014. Deep Domain Confusion: Maximizing for Domain Invariance. arXiv:1412.3474.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, Y.; Chauhan, J.; Wang, W.; and Hsieh, C.-J. 2023. Universality and Limitations of Prompt Tuning. In *Advances in Neural Information Processing Systems*, volume 36, 75623–75643. Curran Associates, Inc.

Wang, Z.; Chi, Z.; Wu, Y.; Gu, L.; Liu, Z.; Plataniotis, K.; and Wang, Y. 2025. Distribution Alignment for Fully Test-Time Adaptation with Dynamic Online Data Streams. In *European Conference on Computer Vision*, 332–349. Cham: Springer Nature Switzerland.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time Adaptation against Multi-modal Reliability Bias. In *International Conference on Learning Representations*.

Yoo, S.; Kim, E.; Jung, D.; Lee, J.; and Yoon, S. 2023. Improving Visual Prompt Tuning for Self-supervised Vision Transformers. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 40075–40092. PMLR.

Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; and Yang, K. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3718–3727. Online: Association for Computational Linguistics.

Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6): 82–88.

Zhang, J.; Huang, J.; Zhang, X.; Shao, L.; and Lu, S. 2024a. Historical Test-time Prompt Tuning for Vision Foundation Models. In *Advances in Neural Information Processing Systems*.

Zhang, M.; Levine, S.; and Finn, C. 2022. MEMO: Test Time Robustness via Adaptation and Augmentation. In *Advances in Neural Information Processing Systems*, volume 35, 38629–38642. Curran Associates, Inc.

Zhang, Y.; and Feng, S. 2023. Enhancing Domain-Invariant Parts for Generalized Zero-Shot Learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 6283–6291. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701085.

Zhang, Z.-Y.; Xie, Z.; Yao, H.; and Sugiyama, M. 2024b. Test-time Adaptation in Non-stationary Environments via Adaptive Representation Alignment. In *Advances in Neural Information Processing Systems*, volume 37, 94607–94632. Curran Associates, Inc.

Zhao, Y.; Luo, J.; Luo, X.; Huang, J.; Yuan, J.; Xiao, Z.; and Zhang, M. 2025. Attention Bootstrapping for Multi-Modal Test-Time Adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 22849–22857.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional Prompt Learning for Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, K.; Yang, Y.; Qiao, Y.; and Xiang, T. 2021. Domain Generalization with MixStyle. In *International Conference on Learning Representations*.