Attention Bootstrapping for Multi-Modal Test-Time Adaptation

Yusheng Zhao[1], Junyu Luo[1], Xiao Luo[2,*], Jinsheng Huang[1],
Jingyang Yuan[1], Zhiping Xiao[3,*], Ming Zhang[1,*]
[1]State Key Laboratory for Multimedia Information Processing,
School of Computer Science, PKU-Anker LLM Lab, Peking University, Beijing, China
[2]Department of Computer Science, University of California, Los Angeles, CA, USA
[3]Paul G. Allen School of Computer Science and Engineering, University of Washington, Seattle, WA, USA
{yusheng.zhao, luojunyu, hjs}@stu.pku.edu.cn, xiaoluo@cs.ucla.edu,
{yuanjy, mzhang_cs}@pku.edu.cn, patxiao@uw.edu

Abstract

Test-time adaptation aims to adapt a well-trained model to potential distribution shifts at test time using only unlabeled test data, without access to the original training data. While previous efforts mainly focus on a single modality, test-time distribution shift in the multi-modal setting is more complex and calls for new solutions. This paper tackles the problem of multi-modal test-time adaptation by proposing a novel method named Attention Bootstrapping with Principal Entropy Minimization (ABPEM). We observe that test-time distribution shift causes misalignment across modalities, leading to a large gap between intra-modality discrepancies (measured by self-attention) and inter-modality discrepancies (measured by cross-attention). We name this the *attention gap*. This attention gap widens with more severe distribution shifts, hindering effective modality fusion. To mitigate this attention gap and encourage better modality fusion, we propose *attention bootstrapping* that promotes cross-attention with the guidance of self-attention. Moreover, to reduce the gradient noise in the commonly-used entropy minimization, we adopt *principal entropy minimization*, a refinement of entropy minimization that reduces gradient noise by focusing on the principal parts of entropy, excluding less reliable gradient information. Extensive experiments on the benchmarks validate the effectiveness of the proposed ABPEM in comparison with competing baselines.

Introduction

Multi-modal learning (Blikstein 2013; Xu, Zhu, and Clifton 2023) has recently attracted increasing attention, with a wide range of applications in many fields, including autonomous driving (Zheng et al. 2023), video understanding (Lee et al. 2023a), sentiment analysis (Yu et al. 2021), and robotics (Krauhausen et al. 2024). Recent advances in this field often encode each modality into tokens and utilize transformers to learn the embedding (Yao and Wan 2020; Zhang et al. 2022). Then, the modalities are often fused with the attention mechanism. Although this paradigm has achieved promising results, it assumes that the test data have the same distribution as the training data, which may fail to hold in the wild (Niu et al. 2023; Tang et al. 2023; Liang, He, and Tan 2024).

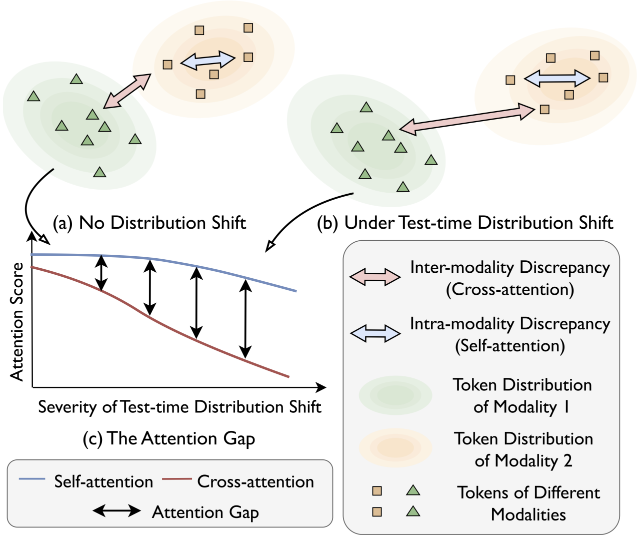

(a) No Distribution Shift
(b) Under Test-time Distribution Shift
Attention Score
Severity of Test-time Distribution Shift
(c) The Attention Gap
Self-attention
Cross-attention
Attention Gap
Inter-modality Discrepancy (Cross-attention)
Intra-modality Discrepancy (Self-attention)
Token Distribution of Modality 1
Token Distribution of Modality 2
Tokens of Different Modalities

Figure 1: During test time, the distribution shift typically has a larger impact on the inter-modality discrepancy than intra-modality discrepancy, leading to an increasing attention gap.

To tackle the challenge of test-time distribution shift, test-time adaptation has emerged as a promising solution as it assumes neither the labels of test data (which is more practical) nor the access of training data (which protects privacy). Recently, many test-time adaptation methods have been proposed (Wang et al. 2021; Boudiaf et al. 2022; Chen et al. 2022; Nguyen et al. 2023; Karmanov et al. 2024). However, most existing test-time adaptation methods focus on the uni-modal setting. In practice, multi-modal test-time adaptation is more challenging. As is shown in Figure 1, the test-time distribution shift causes not only intra-modality changes (blue arrows) but also inter-modality changes (red arrows), and the latter can potentially undermine the model's ability to effectively align and fuse different modalities as the cross-attention tends to decrease under distribution shift.

Towards this end, we propose a novel method named Attention Bootstrapping with Principal Entropy Minimization (ABPEM). [1] As illustrated in Figure 1, when the attention-

[*]Corresponding authors.

[1] In this paper, the term bootstrap is used in its idiomatic sense rather than the statistical sense.

based model is challenged by test-time distribution shift, intra-modality discrepancies increase mildly, while inter-modality discrepancies increase significantly. This is indicated by a mild decrease in the raw self-attention score (before softmax) and a sharp decrease in the raw cross-attention score, which leads to the attention gap. Decreased cross-attention hinders the alignment and fusion across modalities, leading to potentially inferior test-time performance. To encourage cross-attention and decrease inter-modality discrepancies, a naive approach is to minimize the difference between modalities. However, such a method could cause mode collapse and information loss.

To solve this problem, we propose attention bootstrapping that promotes cross-attention scores using self-attention scores. Concretely, we model the distribution of raw cross-/self-attention scores and use the distribution of self-attention scores as the anchor to align the distribution of cross-attention scores. This is better than the naive approach, as it also takes into account the inherent reliability of each modality. It is conceivable that when the intra-modality discrepancies become larger under distribution shift (as indicated by low self-attention scores), this modality becomes less reliable. In such cases, we will have a low anchor (low self-attention scores), and the cross-attention scores have a low target, which results in less attention to this modality.

Moreover, to reduce the noise in the self-supervising signals in multi-modal test-time adaptation, we propose principal entropy minimization. Entropy minimization (Wang et al. 2021; Zhang et al. 2023; Gao, Zhang, and Liu 2024) is a commonly used technique in test-time adaptation in the absence of ground truth labels. However, the computation of entropy involves the model's predictions on every class, reliable ones and unreliable ones. Under test-time distribution shift, the model's predictions on the less-likely classes (classes with lower probabilities) become less reliable, and the gradients of them become noisy. Therefore, the proposed principal entropy minimization excludes the less-likely classes and focuses only on the more-likely (principal) classes, which reduce gradient noise.

Extensive experiments on the benchmarks demonstrate the effectiveness of the proposed method. The contribution of this work is summarized as follows:

- We tackle the problem of multi-modal test-time adaptation, which is practical yet under-explored, and propose a novel method named Attention Bootstrapping with Principal Entropy Minimization (ABPEM).
- We reveal that test-time distribution shift causes modality misalignment, and propose attention bootstrapping to encourage modality alignment and fusion.
- We propose principal entropy minimization that focuses on the principal part of the entropy and reduces the gradient noise in traditional entropy minimization.

Related Works

Test-time Adaptation. Test-time adaptation tackles the problem of distribution shift during test-time, but it assumes neither the knowledge of test data labels nor the access of training data. It is a practical setting since test data labels are hard to obtain, and it also protects privacy (Zhu et al. 2021; Tan et al. 2023). Recently, a number of methods have been proposed to solve this problem, utilizing entropy minimization (Wang et al. 2021; Kundu et al. 2020; Liu, Zhang, and Wang 2021; Mummadi et al. 2021; Lee et al. 2023b, 2024), sample selection (Litrico, Del Bue, and Morerio 2023; Pei et al. 2023), normalization layer tuning (Hu et al. 2021; Yang et al. 2022b; Lim et al. 2023; Wu et al. 2024), representation invariance (Nguyen et al. 2023; Wang et al. 2023; Chen et al. 2023; Ma et al. 2024a), self-supervised learning (Liu et al. 2021; Azimi et al. 2022; Ma 2024), and generative methods (Gao et al. 2023; Prabhudesai et al. 2024; Tsai et al. 2024). While they have achieved remarkable performance, existing efforts mainly focus on a single modality. In multi-modal setting, test-time distribution shift causes not only intra-modal discrepancies but also inter-modal discrepancies. There are some works on multi-modal test-time adaptation, but this work differs from them. Shin et al. (Shin et al. 2022) focus on the specific task of 2D-3D joint segmentation, whereas our method is designed for more general multi-modal settings. Yang et al. (Yang et al. 2024) reveal the challenge of reliability bias caused by multi-modal distribution shifts, and propose READ to tackle the reliability problem. By comparison, we observe a different phenomenon named the attention gap that hinders modality fusion, and propose ABPEM to promote modality fusion under distribution shifts.

Multi-modal Learning. Learning from multi-modal data is an essential topic of deep learning (He et al. 2021; Yang et al. 2022a; Zhao et al. 2022; Ma et al. 2024b; Huang et al. 2024). Recently, there are increasing attention in alignment and fusion of different modalities (Prakash, Chitta, and Geiger 2021; Xu, Yuan, and Ma 2023). Efforts have been devoted to ensure effective fusion in adverse settings, including modality imbalance (Zhou, Chen, and Cao 2020; Peng et al. 2022; Fan et al. 2023), missing modality (Ma et al. 2021, 2022; Woo et al. 2023; Wang et al. 2024), and distribution shift (Liu et al. 2023; Tang et al. 2024; Xia et al. 2024). However, these works focus mainly on the model's training stage, and in resource-limited scenarios, tuning the model's backbone might be infeasible. Moreover, these algorithms often require the labels of the data, which is hard to obtain in practice. This work differs from existing studies, as it explores adapting the model online and during test-time, which is a more practical setting under limited computation resources.

Methodology

Problem Definition

For simplicity, and without loss of generality, we use two modalities (audio, denoted as A, and video, denoted as V) to present the algorithm. The input of each modality is denoted as $\boldsymbol{x}^A$ and $\boldsymbol{x}^V$. The multi-modal learning system encodes the inputs into two sets of tokens in the hidden space, *i.e.* $\{\boldsymbol{z}_i^A\}_{i=1}^{T_A}$ and $\{\boldsymbol{z}_i^V\}_{i=1}^{T_V}$ (where T_A and T_V are the numbers of tokens), using modality-specific encoders, *i.e.* $\mathcal{E}^A$ and $\mathcal{E}^V$. Then, an attention-based fusion module $\mathcal{F}$ is used that combines the two sets of tokens and outputs the probability

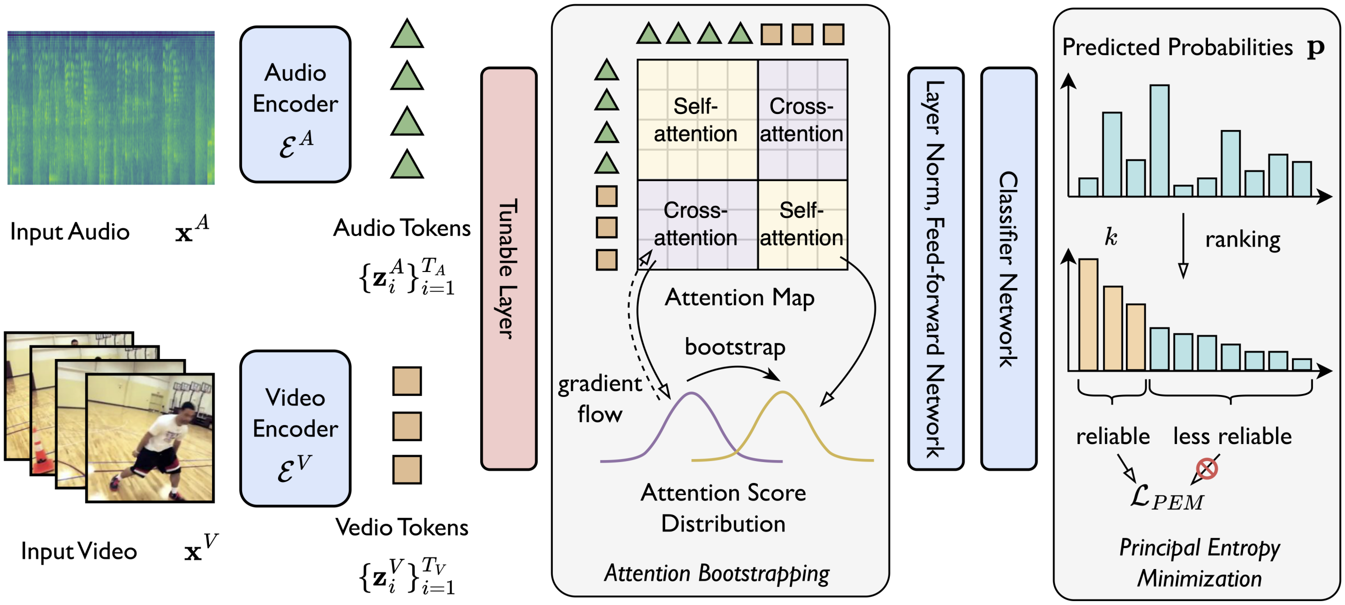


Figure 2: The framework of the proposed ABPEM.

distribution, *i.e.* $\boldsymbol{p} = \mathcal{F}(\{\boldsymbol{z}_i^A\}_{i=1}^{T_A}, \{\boldsymbol{z}_i^V\}_{i=1}^{T_V})$. The probability distribution is denoted as $\boldsymbol{p} = [p_1, \cdots, p_C] \in \Delta^{C-1}$, where C is the number of classes, and Δ^{C-1} is the probability simplex. In multi-modal test-time adaptation, the model $\mathcal{M} = (\mathcal{E}^A, \mathcal{E}^V, \mathcal{F})$ has already been trained on the training set $\mathcal{D}_{tr} = \{(\boldsymbol{x}_i^A, \boldsymbol{x}_i^V, y_i)\}_{i=1}^{N_{tr}}$, where N_{tr} is the size of the training set and y_i is the label. However, the task does not assume access of D_{tr}, and instead, the goal is to improve the model's performance using unlabeled test data $\mathcal{D}_{te} = \{(\boldsymbol{x}_i^A, \boldsymbol{x}_i^V)\}_{i=1}^{N_{te}}$, where N_{te} is the size of the test set. For practicability, we fix $\mathcal{E}^A$, $\mathcal{E}^V$ and only tune a small part of the parameters in $\mathcal{F}$.

Framework Overview

The framework of the proposed ABPEM is illustrated in Figure 2. During test time, the input audio and video are encoded by $\mathcal{E}^A$ and $\mathcal{E}^B$ to obtain hidden space representations (*i.e.* tokens). Then, the tokens from different modalities are concatenated and a tunable layer is applied to generate queries, keys and values for each token, which are used for attention. The attention bootstrapping is used on the attention map that aligns the distributions of cross- and self-attention. Subsequently, the attended tokens are processed by layer normalization, the feed-forward network, and the classifier network to generate predicted probabilities $\boldsymbol{p} \in \Delta^{C-1}$. Finally, principal entropy minimization takes top k reliable classes and computes the principal part of the entropy, which is part of the objective.

Attention Bootstrapping

The attention mechanism (Vaswani et al. 2017) is the most widely used paradigm for modality fusion (Nagrani et al. 2021; Zong and Sun 2023). However, when the model is challenged by multi-modal test-time distribution shift, inter-modality discrepancy experiences more increase than intra-modality discrepancy, which leads to a widening gap between self-attention and cross-attention. Attention bootstrapping aims to bootstrap cross-attention using self-attention. Under the aforementioned paradigm, the tokens learned by the modality-specific encoders are first concatenated as $\boldsymbol{Z} = [\boldsymbol{z}_1^A, \cdots, \boldsymbol{z}_{T_A}^A, \boldsymbol{z}_1^V, \cdots, \boldsymbol{z}_{T_V}^V]^T$. Subsequently, query, key and value matrices $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})$ are computed as:

$$\begin{aligned} \boldsymbol{Q} = \boldsymbol{Z}\boldsymbol{W}_Q + \boldsymbol{B}_Q, \quad \boldsymbol{K} = \boldsymbol{Z}\boldsymbol{W}_K + \boldsymbol{B}_K, \\ \boldsymbol{V} = \boldsymbol{Z}\boldsymbol{W}_V + \boldsymbol{B}_V, \end{aligned} \tag{1}$$

where $\boldsymbol{W}_{Q,K,V}$ and $\boldsymbol{B}_{Q,K,V}$ are learnable parameters during test time. Then, the unnormalized attention can be computed as follows using queries and keys:

$$\tilde{\boldsymbol{A}} = \boldsymbol{Q}\boldsymbol{K}^T. \tag{2}$$

We decompose $\tilde{\boldsymbol{A}}$ into four parts, *i.e.*

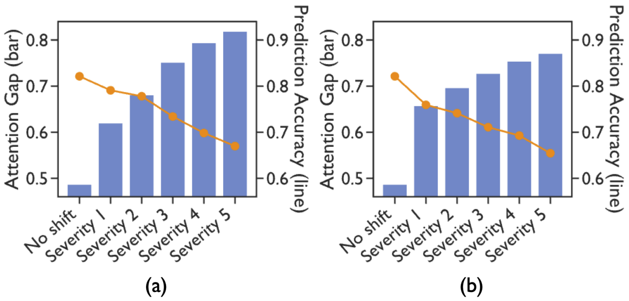
$$\tilde{\boldsymbol{A}} = \begin{bmatrix} \tilde{\boldsymbol{A}}^{A2A} & \tilde{\boldsymbol{A}}^{A2V} \\ \tilde{\boldsymbol{A}}^{V2A} & \tilde{\boldsymbol{A}}^{V2V} \end{bmatrix}, \tag{3}$$

where the $X2Y$ superscript denotes the unnormalized attention when modality X is the query and modality Y is the key. The normalized version of attention scores are used to fuse different modalities:

$$\boldsymbol{A} = \mathrm{softmax}(\tilde{\boldsymbol{A}}/\sqrt{d}), \quad \boldsymbol{Z}' = \boldsymbol{A}\boldsymbol{V}, \tag{4}$$

where d is the dimension of tokens and $\boldsymbol{Z}'$ is the attended token embeddings.

When the model experiences test-time distribution shifts, modality mismatch may happen, and this will cause both intra- and inter- modality discrepancy. For example, when the input video is subject to distribution shift, the distribution of token embeddings $\{\boldsymbol{z}_i^V\}_{i=1}^{T_V}$ will have slightly higher variance due to the increased uncertainty (intra-modality discrepancy), and they will also shift away from audio tokens



(a)

(b)

Figure 3: As the test-time distribution shift becomes severer, the attention gap (blue bar plot) tends to increase, and the prediction accuracy (orange line plot) tends to decrease.

(inter-modality discrepancy). The intra-modality discrepancy is signified by the decreasing value of self-attention scores (*i.e.* $\tilde{\boldsymbol{A}}^{V2V}$), and inter-modality discrepancy is shown by the decreasing of cross-attention (*i.e.* $\tilde{\boldsymbol{A}}^{A2V}$). Note that we use the unnormalized attention scores (the ones before softmax), as they better reflect the distance of distributions. Normalized attention scores are influenced by many factors (*e.g.* the number of tokens). To better describe the attention scores, we model them as Gaussian distributions:

$$\begin{gathered} P_{A2A}(a) \sim \mathcal{N}(\mu_{A2A}; \sigma^2_{A2A}), \\ \mu_{A2A} = \mathrm{avg}(\tilde{\boldsymbol{A}}^{A2A}_{ij}), \quad \sigma^2_{A2A} = \mathrm{var}(\tilde{\boldsymbol{A}}^{A2A}_{ij}), \end{gathered} \tag{5}$$

where $i = 1, 2, \cdots, T_A$ and $j = 1, 2, \cdots, T_A$. Similarly, $P_{A2V}(a)$, $P_{V2V}(a)$, and $P_{V2A}(a)$ can be defined.

In Figure 3, we provide the empirical evidence that the attention gap exists and tends to increase as the test-time distribution shift becomes severer. Specifically, we introduce two types of noise to the vision modality, *i.e.* defocus blur (a) and zoom blur (b), and measure the attention gap $\mu_{V2V} - \mu_{A2V}$ as well as the prediction accuracy of the model. As can be seen from the figure, when the model faces test-time distribution shift, the attention gap (blue bar plot) generally increases and the prediction accuracy (orange line plot) drops. This suggests that inter-modality dependencies (signified by cross-attention scores, *i.e.* $P_{A2V}(a)$ and $P_{V2A}(a)$) are more affected than intra-modality dependencies (signified by self-attention scores, *i.e.* $P_{A2A}(a)$ and $P_{V2V}(a)$) under distribution shift. Therefore, a feasible solution is to perform attention bootstrapping that uses self-attention scores as anchors to boost cross-attention scores, and thus reduce the attention gap, encouraging modality fusion.

Specifically, we adopt the strategy that minimizes the Kullback-Leibler divergence between the distributions of attention scores, which is formulated as follows:

$$\begin{aligned} \mathcal{D}_{KL}(P_{A2V} || P_{V2V}) = \log \frac{\sigma_{V2V}}{\sigma_{A2V}} - \frac{1}{2} \\ + \frac{\sigma^2_{A2V} + (\mu_{A2V} - \mu_{V2V})^2}{2\sigma^2_{V2V}}. \end{aligned} \tag{6}$$

In Eq. 6, P_{V2V} reflects how the video modality attends itself, and P_{A2V} describes how the audio modality attends the video modality. In other words, P_{V2V} reflects the video's evaluation of itself: μ_{V2V} is an evaluation of the amount of information relevant to the prediction task, while σ_{V2V} is an evaluation of discriminability across tokens. When the video modality itself has been fully adapted to the distribution shift, it is conceivable that such evaluation is better than the assessment from the audio modality (μ_{A2V} and σ_{A2V}) since there are modality discrepancies. Thus, our goal is to decrease inter-modality discrepancies so that they are similar to intra-modality ones (not decreased to zero as we want to preserve the natural discrepancies of different tokens). Therefore, it is reasonable to use P_{V2V} as an anchor to bootstrap P_{A2V}. We stop the gradient flow of the self-attention scores μ_{V2V} and σ_{V2V} to avoid influencing the anchor.

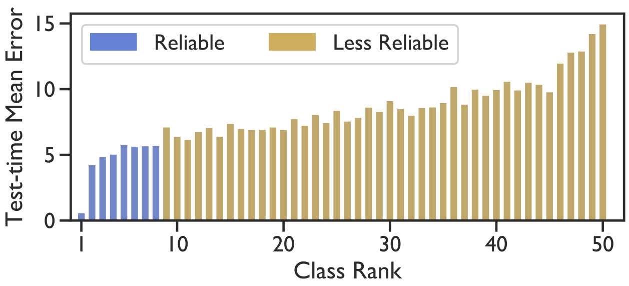


Figure 4: The test-time mean error increases with the rank of the class. Classes with lower ranks are more robust to test-time distribution shift (lower errors).

Similarly, $\mathcal{D}_{KL}(P_{V2A} || P_{A2A})$ can be calculated, and the loss objective of attention bootstrapping is written as:

$$\mathcal{L}_{AB} = \mathcal{D}_{KL}(P_{A2V} || P_{V2V}) + \mathcal{D}_{KL}(P_{V2A} || P_{A2A}). \tag{7}$$

Principal Entropy Minimization

Entropy minimization is a commonly used technique in test-time adaptation (Wang et al. 2021; Liu, Zhang, and Wang 2021; Lee et al. 2023b), as it does not require the labels to compute loss objective. Although it improves performance, entropy minimization inevitably introduce noisy gradient signals, and there are some works that tackle this problem from the sample perspective (Zhang et al. 2023; Gao, Zhang, and Liu 2024; Xiong and Xiang 2024). However, these methods may fail to utilize all the samples from the test data. In this part, we introduce principal entropy minimization that tackles this problem from the class perspective. Specifically, we can write the entropy of $\boldsymbol{p}$ as:

$$\mathcal{H}(\boldsymbol{p}) = -\sum_{i \in \mathcal{S}} p_i \log p_i, \tag{8}$$

where $\mathcal{S} = \{1, 2, \cdots, C\}$ is the set of all classes.

However, Eq. 8 contains terms of all classes, including the more reliable ones and the less reliable ones. Denote the rank of each class as $r_i, i \in \mathcal{S}$, which can be formally defined as:

$$r_i = |\{p_j \mid j \in \mathcal{S} \wedge p_j \geq p_i\}|, \tag{9}$$

where $|\cdot|$ denotes the cardinality of a set. Our observation is that classes with lower ranks are more reliable. The empirical evidence is presented in Figure 4, where we measure the

Algorithm 1: Optimization Algorithm of ABPEM

```
Requires: The well-trained model $\mathcal{M} = (\mathcal{E}^A, \mathcal{E}^V, \mathcal{F})$, the unlabeled test dataset $\mathcal{D}_{te}$, and $k$ in Eq. 10.
Ensures: The adapted model, and the prediction on test data.
for each batch in $\mathcal{D}_{te}$ do
    Compute the embeddings of inputs, i.e. $\{\boldsymbol{z}_i^A\}_{i=1}^{T_A} = \mathcal{E}^A(\boldsymbol{x}^A)$ and $\{\boldsymbol{z}_i^V\}_{i=1}^{T_V} = \mathcal{E}^V(\boldsymbol{x}^V)$
    Compute the attention scores using Eq. 2.
    Compute the attention bootstrapping loss using Eq. 7.
    Obtain the predicted probabilities $\boldsymbol{p}$ from the output of the fusion module $\mathcal{F}$.
    Sort the predicted probabilities to obtain the ranks $r_i$ defined in Eq. 10.
    Compute the principal entropy in Eq. 11 as $\mathcal{L}_{PEM}$.
    Compute the final loss function in Eq. 13.
    Update the tunable parameters in the fusion module $\mathcal{F}$ through back-propagation.
end for
```

changes of predicted probability with respect to the rank of the class. The results show that classes with lower ranks (or relatively higher probabilities) are more robust to test-time distribution shifts. Therefore, it is reasonable to exclude the less reliable set from $\mathcal{S}$ in the computation of entropy.

Specifically, we define the reliable class set $\mathcal{S}_R^{(k)}$ *for each test sample* based on the ranks r_i derived from the predicted probabilities p_i, which is defined as follows:

$$\mathcal{S}_R^{(k)} = \{i \in \mathcal{S} | r_i \leq k\}, \tag{10}$$

where k is a hyper-parameter. Subsequently, we define the principal entropy of $\boldsymbol{p}$ as:

$$\mathcal{H}_P^{(k)}(\boldsymbol{p}) = -\sum_{i \in \mathcal{S}_R^{(k)}} p_i \log p_i. \tag{11}$$

The principal entropy is then used as the minimizing objective to replace the entropy defined in Eq. 8:

$$\mathcal{L}_{PEM} = \mathcal{H}_P^{(k)}(\boldsymbol{p}). \tag{12}$$

Summary

In this part, we provide a summary of our method. When the multi-modal learning system $\mathcal{M}$ receives data, it first encodes the inputs of each modality using modality-specific encoders $\mathcal{E}^A$ and $\mathcal{E}^V$. Then the learned embeddings are sent into the fusion module $\mathcal{F}$, in which attention bootstrapping is performed using the attention scores. The fusion module yields a probability distribution for each sample, and the principal entropy is computed as the loss objective. The final loss function can be written as:

$$\mathcal{L} = \lambda \mathcal{L}_{AB} + \mathcal{L}_{PEM} \tag{13}$$

The optimization procedure is summarized in Algorithm 1. It can be shown that this algorithm has the same time complexity as the model $\mathcal{M}$ without adaptation, which is $\mathcal{O}(N_{te} d (T_A + T_V)^2)$. Empirical results about efficiency are provided in the experiment section.

Experiments

Experimental Settings

Benchmarks. The experiments are performed on two benchmarks: Kinetics50-C and VGGSound-C (Yang et al. 2024), which are based on the widely used Kinetics (Kay et al. 2017) and VGGSound (Chen et al. 2020) datasets. Each of the benchmarks contains two settings, *i.e.* corrupted video setting (which contains 15 types of video corruptions) and corrupted audio setting (which contains 6 types of audio corruptions). Each type of corruption has 5 severity, and we adopt severity 5 as default following (Yang et al. 2024). In the Kinetics50 dataset (from which Kinetics50-C benchmark is constructed), the video modality typically contains more information, whereas in the VGGSound dataset (from which VGGSound-C is constructed), the audio modality typically contains more information.

Baselines Methods. The proposed ABPEM is compared with several competing baselines, including Tent (Wang et al. 2021), MMT (Shin et al. 2022), EATA (Niu et al. 2022), SAR (Niu et al. 2023), and READ (Yang et al. 2024).

Implementation Details. In the experiments, we use CAV-MAE (Gong et al. 2023) as the architecture of $\mathcal{M}$. The model is pretrained on the corresponding training set (Kinetics or VGGSound). We set k in Eq. 10 to about 8 for Kinetics50-C and 30 for VGGSound-C, and λ to 1 by default. Moreover, we also use a class-balancing loss in alignment with (Yang et al. 2024). For optimization, we use Adam optimizer (Kingma and Ba 2014) and the model is optimized within a single epoch, with the learning rate of 1×10^{-4}. More details can be found at https://github.com/YushengZhao/ABPEM.

Performance Comparison

We first compare the performance of the proposed ABPEM and the baselines in Table 1, Table 2 and Table 3. The first model (Raw) denotes the model without any test-time adaptation. From the results, we have several observations.

- The proposed ABPEM achieves a consistent lead in both Kinetics50-C and VGGSound-C benchmarks in the face of various types of test-time distribution shifts. This shows the overall effectiveness of the proposed ABPEM.
- Our model experiences more significant improvement when facing the test-time distribution that affects the more informative modality (*i.e.* corrupted video modality of Kinetics50-C and corrupted audio modality of VGGSound). Previous efforts often ignore the problem of increasing attention gap, whereas the proposed ABPEM explicitly reduces this gap, which is beneficial for the model to incorporate the more informative modality with the other modality, and thus achieves higher accuracy.
- The task of multi-modal test-time adaptation is inherently hard. When the model is challenged by adverse distribution shifts, some models fail to achieve satisfactory performance, and sometimes even worse than the model before adaptation. This shows that without ground truth labels, the gradients can be very noisy and are potentially

Models	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	
Raw	46.8	48.0	46.9	67.5	62.2	70.6	67.7	61.6	60.3	46.7	75.2	52.1	65.7	66.5	61.9	59.9
MMT	46.2	46.6	46.1	58.8	55.7	62.4	61.7	52.6	54.4	48.5	69.3	49.3	57.6	56.4	54.5	54.5
Tent	46.3	47.0	46.3	67.4	62.5	70.4	67.7	63.1	61.1	34.9	75.4	51.6	66.7	66.5	62.0	59.4
EATA	46.8	47.6	47.1	67.2	61.8	70.2	67.7	61.6	60.6	46.0	75.2	52.4	65.9	66.4	62.7	60.1
SAR	46.7	47.4	46.6	67.0	61.7	70.0	66.4	61.8	60.6	46.0	75.2	52.1	65.7	66.0	62.0	59.8
READ	49.4	49.7	49.0	68.0	65.1	71.2	69.0	64.5	64.4	57.4	75.5	53.6	68.3	68.0	65.1	62.5
ABPEM	**50.3**	**51.1**	**50.4**	**70.0**	**69.6**	**72.5**	**71.2**	**65.2**	**66.2**	**65.6**	**75.7**	**56.6**	**71.9**	**70.5**	**67.8**	**65.0**

Table 1: Prediction accuracies (in %) on Kinetics50-C benchmark (corrupted video modality).

Models	Noise			Weather			Avg.	Noise			Weather			Avg.
	Gauss.	Traff.	Crowd.	Rain	Thund.	Wind		Gauss.	Traff.	Crowd.	Rain	Thund.	Wind	
Raw	73.7	65.5	67.9	70.3	67.9	70.3	69.3	37.0	25.5	16.8	21.6	27.3	25.5	25.6
MMT	70.8	69.2	68.5	69.0	69.8	68.5	69.4	14.1	5.2	6.4	9.8	8.6	4.5	7.6
Tent	73.9	67.4	68.5	70.4	66.5	70.4	69.6	10.6	2.6	1.8	2.3	3.3	4.1	4.5
EATA	73.7	66.1	68.5	69.5	70.6	69.4	69.4	39.2	26.1	22.9	26.0	31.7	30.4	29.4
SAR	73.7	65.4	68.2	69.9	67.2	70.2	69.1	37.4	9.5	11.0	12.1	26.8	23.7	20.1
READ	74.1	69.0	69.7	71.1	71.8	70.7	71.1	40.4	28.9	26.6	30.9	36.7	30.6	32.4
ABPEM	**74.8**	**71.3**	**71.5**	**71.9**	**73.8**	**71.6**	**72.5**	**40.6**	**33.7**	**34.8**	**32.2**	**41.1**	**34.4**	**36.1**

Table 2: Prediction accuracies (in %) on Kinetics50-C (left) and VGGSound- C (right) benchmarks (corrupted audio modality).

Models	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impul.	Defoc.	Glass	Mot.	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elas.	Pix.	JPEG	
Raw	52.8	52.7	52.7	57.2	57.2	58.7	56.8	56.4	56.6	55.6	58.9	53.7	56.9	55.8	56.9	56.0
MMT	7.1	7.3	7.3	44.8	41.5	48.0	45.5	27.4	23.5	30.5	46.3	24.0	43.0	40.7	45.7	32.0
Tent	52.7	52.7	52.7	56.7	56.5	58.0	56.5	55.0	57.0	56.3	58.7	54.0	57.4	56.7	57.4	55.8
EATA	53.0	52.8	53.0	57.2	57.1	58.6	57.8	56.3	56.8	56.4	59.0	54.1	57.4	56.1	57.0	56.2
SAR	52.9	52.8	52.9	57.0	57.1	58.5	56.8	56.3	56.7	55.9	58.9	54.0	57.6	57.1	57.2	56.1
READ	53.6	53.6	53.5	57.9	57.7	59.4	58.8	57.2	57.8	55.0	59.9	55.2	58.6	57.1	57.9	56.9
ABPEM	**54.0**	**53.9**	**54.0**	**58.2**	**58.1**	**59.6**	**59.3**	**57.5**	**58.2**	**58.2**	**60.2**	**56.2**	**59.1**	**57.5**	**58.3**	**57.5**

Table 3: Prediction accuracies (in %) on VGGSound-C benchmark (corrupted video modality).

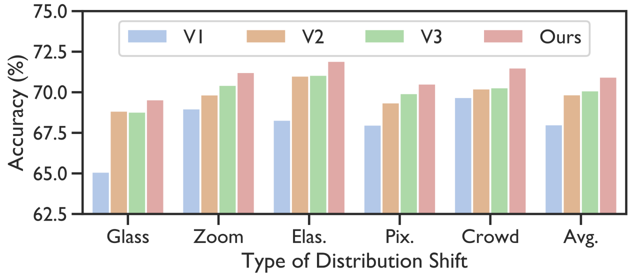


Figure 5: Ablation of the main components of ABPEM.

harmful for the model. Our model adopts principal entropy minimization, which reduces the noise in the gradient and leads to better results.

Ablation Studies

Ablation of the main components. We design several variants of the model to investigate the role of attention bootstrapping and principal entropy minimization. V1 is the model that adopt a tunable layer , and use a basic self-supervised objective in (Yang et al. 2024). V2 is the model that uses attention bootstrapping. V3 is the model that replaces principal entropy minimization with vanilla entropy minimization. The last model is the proposed ABPEM, which contains both attention bootstrapping and principal entropy minimization. The results on Kinetics50-C benchmark are shown in Figure 5. As can be seen from the results, the use of attention bootstrapping increases accuracy significantly (comparing V1 and V2), this can be attributed to the better fusion of modalities under distribution shift. Moreover, the improvement of vanilla entropy minimization is marginal (comparing V2 and V3), whereas the proposed principal entropy minimization further boosts the accuracy (comparing V2 and Ours). This shows that reducing the noise in the gradients, which is achieved by principal entropy minimization, is beneficial for the performance.

Ablation of k in Eq. 10. We then investigate the role of hyperparameter k in principal entropy minimization. The results on Kinetics50-C benchmark are shown in Table 5. As can be seen from the table, the model is generally not sensitive to k, and the highest accuracy is achieved at 8. When k is small, the reliable class set $\mathcal{S}_R^{(k)}$ for each sample is small, which may not fully utilize all the information. Conversely, when k is large, the reliability of class probabilities p_i decreases, leading to noisy gradient information.

Models	Raw	Tent	EATA	SAR	READ	ABPEM
Samples per second	92.4	68.5	69.8	55.6	88.2	87.3
# Tunable parameters	0	0.2M	0.2M	0.2M	1.8M	1.8M

Table 4: Comparison of models' efficiency.

k	Glass	Zoom	Elas.	Pix.	Crowd	Avg.
6	69.28	70.83	71.71	69.88	71.49	70.64
7	69.40	71.18	71.89	70.42	71.48	70.87
8	69.56	71.24	71.93	70.53	71.52	70.96
9	69.53	71.33	71.83	70.59	71.49	70.95
10	69.55	71.34	71.75	70.61	71.48	70.94

Table 5: Ablated study about k in Eq. 10.

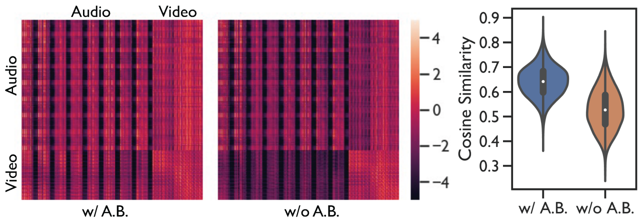


Figure 6: The attention map with attention bootstrapping (w/ A.B., left) and without (w/o A.B., middle). The proposed attention bootstrapping encourages cross-attention, and achieves better alignment with increased cosine similarity between token embeddings of different modalities.

Efficiency Comparison

We then show that the proposed ABPEM is efficient. As mentioned in the previous section, the time complexity of the model is equivalent to the raw model $\mathcal{M}$ without adaptation. In Table 4, empirical evidences are provided where we count the number of samples that the model processes per second and the total number of tunable parameters on the Kinetics50-C dataset. As is shown in the table, ABPEM achieves similar speed (samples per second) compared to READ. Moreover, it is faster than methods that require tuning layer normalization modules (Tent, EATA, and SAR) with relatively more tunable parameters, similar to READ. The results show the efficiency of the proposed method.

Further Analysis

Better alignment with decreased attention gap. We investigate the role of attention bootstrapping in the alignment and fusion of different modalities. As mentioned before, attention bootstrapping encourages cross-attention with the help of self-attention, and we visualize the unnormalized attention map $\tilde{\boldsymbol{A}}$ of Eq. 2 in Figure 6 (left and middle, on the Kinetics50-C benchmark, Fog corruption). The results show a significant increase in cross-attention scores, and the attention gap $\mu_{V2V} - \mu_{A2V}$ reduces from 1.02 to 0.11 (another gap $\mu_{A2A} - \mu_{V2A}$ reduces from 0.98 to 0.23). This leads to more aligned representations, as shown in Figure 6 (right), where we plot the distribution of cosine similarities between the token embeddings of two modalities. This shows that the representations of different modalities are more aligned (higher similarity) when attention bootstrapping is used.

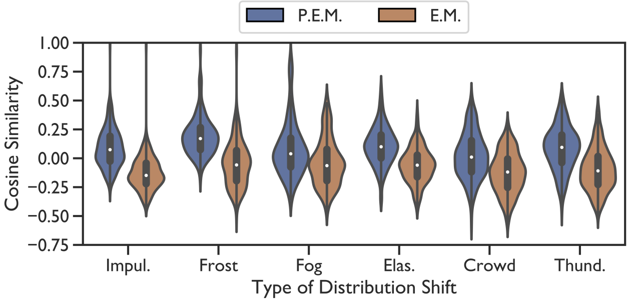


Figure 7: The cosine similarities between the gradients of cross entropy objective (with ground truth labels) and the gradients of principal entropy minimization (P.E.M.) / entropy minimization (E.M.) objective. P.E.M. yields gradient directions closer to the ground truth than E.M.

Reduced gradient noise. We then show the effect of principal entropy minimization in the reduction of gradient noise. Specifically, we perform experiments on Kinetics50-C under various types of distribution shifts and compare the gradients generated from two sources: (1) the cross entropy loss objective using the ground truth labels of the samples, and (2) the principal entropy minimization (P.E.M.) objective or the commonly used entropy minimization (E.M.) objective. The distributions of the cosine similarities between gradients from sources (1) and (2) are illustrated in Figure 7. As can be seen from the figure, the gradients generated by P.E.M. objective are closer to the gradients generated using ground truth labels. This shows that by excluding the less reliable classes from the computation of entropy, P.E.M. objective reduces the noise signals in the gradients.

Conclusion

This paper tackles the problem of multi-modal test-time adaptation, and proposes attention bootstrapping and principal entropy minimization (ABPEM) to solve this problem. When the multi-modal learning system is influenced by distribution shifts, modality mismatch occurs, and the attention gap increases, which hinders the fusion of different modalities. To reduce this gap, attention bootstrapping is proposed. Moreover, we observe that classes with lower probabilities are less reliable and may introduce noise in the gradients. To tackle this, we differentiate the reliable classes and less reliable classes with the proposed principal entropy minimization. Extensive experiments on the benchmark datasets demonstrate the effectiveness of the proposed ABPEM.

Acknowledgments

This paper is partially supported by the National Key Research and Development Program of China with Grant No. 2023YFC3341203 as well as the National Natural Science Foundation of China with Grant Number 62276002.

References

Azimi, F.; Palacio, S.; Raue, F.; Hees, J.; Bertinetto, L.; and Dengel, A. 2022. Self-supervised test-time adaptation on video data. In *WACV*, 3439–3448.

Blikstein, P. 2013. Multimodal learning analytics. In *LAK*, 102–106.

Boudiaf, M.; Mueller, R.; Ben Ayed, I.; and Bertinetto, L. 2022. Parameter-free online test-time adaptation. In *CVPR*, 8344–8353.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022. Contrastive test-time adaptation. In *CVPR*, 295–305.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 721–725. IEEE.

Chen, L.; Zhang, Y.; Song, Y.; Shan, Y.; and Liu, L. 2023. Improved test-time adaptation for domain generalization. In *CVPR*, 24172–24182.

Fan, Y.; Xu, W.; Wang, H.; Wang, J.; and Guo, S. 2023. Pmr: Prototypical modal rebalance for multimodal learning. In *CVPR*, 20029–20038.

Gao, J.; Zhang, J.; Liu, X.; Darrell, T.; Shelhamer, E.; and Wang, D. 2023. Back to the source: Diffusion-driven adaptation to test-time corruption. In *CVPR*, 11786–11796.

Gao, Z.; Zhang, X.-Y.; and Liu, C.-L. 2024. Unified Entropy Optimization for Open-Set Test-Time Adaptation. In *CVPR*, 23975–23984.

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

He, D.; Zhao, Y.; Luo, J.; Hui, T.; Huang, S.; Zhang, A.; and Liu, S. 2021. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *ACM MM*, 2344–2352.

Hu, X.; Uzunbas, G.; Chen, S.; Wang, R.; Shah, A.; Nevatia, R.; and Lim, S.-N. 2021. Mixnorm: Test-time adaptation through online normalization estimation. *arXiv preprint arXiv:2110.11478*.

Huang, J.; Chen, L.; Guo, T.; Zeng, F.; Zhao, Y.; Wu, B.; Yuan, Y.; Zhao, H.; Guo, Z.; Zhang, Y.; et al. 2024. Mmevalpro: Calibrating multimodal benchmarks towards trustworthy and efficient evaluation. *arXiv preprint arXiv:2407.00468*.

Karmanov, A.; Guan, D.; Lu, S.; El Saddik, A.; and Xing, E. 2024. Efficient Test-Time Adaptation of Vision-Language Models. In *CVPR*, 14162–14171.

Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krauhausen, I.; Griggs, S.; McCulloch, I.; den Toonder, J. M.; Gkoupidenis, P.; and van de Burgt, Y. 2024. Bio-inspired multimodal learning with organic neuromorphic electronics for behavioral conditioning in robotics. *Nature Communications*, 15(1): 4765.

Kundu, J. N.; Venkat, N.; Babu, R. V.; et al. 2020. Universal source-free domain adaptation. In *CVPR*, 4544–4553.

Lee, D. W.; Ahuja, C.; Liang, P. P.; Natu, S.; and Morency, L.-P. 2023a. Lecture presentations multimodal dataset: Towards understanding multimodality in educational videos. In *ICCV*, 20087–20098.

Lee, J.; Das, D.; Choo, J.; and Choi, S. 2023b. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *ICCV*, 16380–16389.

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. *arXiv preprint arXiv:2403.07366*.

Liang, J.; He, R.; and Tan, T. 2024. A comprehensive survey on test-time adaptation under distribution shifts. *IJCV*, 1–34.

Lim, H.; Kim, B.; Choo, J.; and Choi, S. 2023. TTN: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*.

Litrico, M.; Del Bue, A.; and Morerio, P. 2023. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *CVPR*, 7640–7650.

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? *NeurIPS*, 34: 21808–21820.

Liu, Y.; Qiao, L.; Lu, C.; Yin, D.; Lin, C.; Peng, H.; and Ren, B. 2023. OSAN: A one-stage alignment network to unify multimodal alignment and unsupervised domain adaptation. In *CVPR*, 3551–3560.

Liu, Y.; Zhang, W.; and Wang, J. 2021. Source-free domain adaptation for semantic segmentation. In *CVPR*, 1215–1224.

Ma, H.; Zhu, Y.; Zhang, C.; Zhao, P.; Wu, B.; Huang, L.-K.; Hu, Q.; and Wu, B. 2024a. Invariant Test-Time Adaptation for Vision-Language Model Generalization. *arXiv preprint arXiv:2403.00376*.

Ma, J. 2024. Improved Self-Training for Test-Time Adaptation. In *CVPR*, 23701–23710.

Ma, M.; Ren, J.; Zhao, L.; Testuggine, D.; and Peng, X. 2022. Are multimodal transformers robust to missing modality? In *CVPR*, 18177–18186.

Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; and Peng, X. 2021. Smil: Multimodal learning with severely missing modality. In *AAAI*, volume 35, 2302–2310.

Ma, X.; Yang, M.; Li, Y.; Hu, P.; Lv, J.; and Peng, X. 2024b. Cross-modal Retrieval with Noisy Correspondence via Consistency Refining and Mining. *TIP*.

Mummadi, C. K.; Hutmacher, R.; Rambach, K.; Levinkov, E.; Brox, T.; and Metzen, J. H. 2021. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*.

Nagrani, A.; Yang, S.; Arnab, A.; Jansen, A.; Schmid, C.; and Sun, C. 2021. Attention bottlenecks for multimodal fusion. *NeurIPS*, 34: 14200–14213.

Nguyen, A. T.; Nguyen-Tang, T.; Lim, S.-N.; and Torr, P. H. 2023. Tipi: Test time adaptation with transformation invariance. In *CVPR*, 24162–24171.

Niu, S.; Wu, J.; Zhang, Y.; Chen, Y.; Zheng, S.; Zhao, P.; and Tan, M. 2022. Efficient test-time model adaptation without forgetting. In *ICML*, 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. *arXiv preprint arXiv:2302.12400*.

Pei, J.; Jiang, Z.; Men, A.; Chen, L.; Liu, Y.; and Chen, Q. 2023. Uncertainty-induced transferability representation for source-free unsupervised domain adaptation. *TIP*, 32: 2033–2048.

Peng, X.; Wei, Y.; Deng, A.; Wang, D.; and Hu, D. 2022. Balanced multimodal learning via on-the-fly gradient modulation. In *CVPR*, 8238–8247.

Prabhudesai, M.; Ke, T.-W.; Li, A.; Pathak, D.; and Fragkiadaki, K. 2024. Test-time adaptation of discriminative models via diffusion generative feedback. *NeurIPS*, 36.

Prakash, A.; Chitta, K.; and Geiger, A. 2021. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 7077–7087.

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 *CVPR*, 16928–16937.

Tan, Y.; Liu, Y.; Long, G.; Jiang, J.; Lu, Q.; and Zhang, C. 2023. Federated learning on non-iid graphs via structural knowledge sharing. In *AAAI*, volume 37, 9953–9961.

Tang, J.; Chen, S.; Niu, G.; Sugiyama, M.; and Gong, C. 2023. Distribution shift matters for knowledge distillation with webly collected images. In *ICCV*, 17470–17480.

Tang, S.; Su, W.; Ye, M.; and Zhu, X. 2024. Source-Free Domain Adaptation with Frozen Multimodal Foundation Model. In *CVPR*, 23711–23720.

Tsai, Y.-Y.; Chen, F.-C.; Chen, A. Y.; Yang, J.; Su, C.-C.; Sun, M.; and Kuo, C.-H. 2024. GDA: Generalized Diffusion for Robust Test-time Adaptation. In *CVPR*, 23242–23251.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*, 30.

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

Wang, H.; Luo, S.; Hu, G.; and Zhang, J. 2024. Gradient-Guided Modality Decoupling for Missing-Modality Robustness. In *AAAI*, volume 38, 15483–15491.

Wang, S.; Zhang, D.; Yan, Z.; Zhang, J.; and Li, R. 2023. Feature alignment and uniformity for test time adaptation. In *CVPR*, 20050–20060.

Woo, S.; Lee, S.; Park, Y.; Nugroho, M. A.; and Kim, C. 2023. Towards good practices for missing modality robust action recognition. In *AAAI*, volume 37, 2776–2784.

Wu, Y.; Chi, Z.; Wang, Y.; Plataniotis, K. N.; and Feng, S. 2024. Test-time domain adaptation by learning domain-aware batch normalization. In *AAAI*, volume 38, 15961–15969.

Xia, Y.; Huang, H.; Zhu, J.; and Zhao, Z. 2024. Achieving cross modal generalization with multimodal unified representation. *NeurIPS*, 36.

Xiong, H.; and Xiang, Y. 2024. Robust gradient aware and reliable entropy minimization for stable test-time adaptation in dynamic scenarios. *The Visual Computer*, 1–16.

Xu, H.; Yuan, J.; and Ma, J. 2023. Murf: Mutually reinforcing multi-modal image registration and fusion. *TPAMI*, 45(10): 12148–12166.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal learning with transformers: A survey. *TPAMI*, 45(10): 12113–12132.

Yang, M.; Huang, Z.; Hu, P.; Li, T.; Lv, J.; and Peng, X. 2022a. Learning with twin noisy labels for visible-infrared person re-identification. In *CVPR*, 14308–14317.

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

Yang, T.; Zhou, S.; Wang, Y.; Lu, Y.; and Zheng, N. 2022b. Test-time batch normalization. *arXiv preprint arXiv:2205.10210*.

Yao, S.; and Wan, X. 2020. Multimodal transformer for multimodal machine translation. In *ACL*, 4346–4350.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, volume 35, 10790–10797.

Zhang, J.; Qi, L.; Shi, Y.; and Gao, Y. 2023. Domainadaptor: A novel approach to test-time adaptation. In *ICCV*, 18971–18981.

Zhang, W.; Qiu, F.; Wang, S.; Zeng, H.; Zhang, Z.; An, R.; Ma, B.; and Ding, Y. 2022. Transformer-based multimodal information fusion for facial expression analysis. In *CVPR*, 2428–2437.

Zhao, Y.; Chen, J.; Gao, C.; Wang, W.; Yang, L.; Ren, H.; Xia, H.; and Liu, S. 2022. Target-driven structured transformer planner for vision-language navigation. In *ACM MM*, 4194–4203.

Zheng, T.; Li, A.; Chen, Z.; Wang, H.; and Luo, J. 2023. Autofed: Heterogeneity-aware federated multimodal learning for robust autonomous driving. In *MobiCom*, 1–15.

Zhou, K.; Chen, L.; and Cao, X. 2020. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, 787–803. Springer.

Zhu, H.; Xu, J.; Liu, S.; and Jin, Y. 2021. Federated learning on non-IID data: A survey. *Neurocomputing*, 465: 371–390.

Zong, D.; and Sun, S. 2023. Mcomet: Multimodal fusion transformer for physical audiovisual commonsense reasoning. In *AAAI*, volume 37, 6621–6629.