Bridging the Gap for Test-Time Multimodal Sentiment Analysis

Zirun Guo*, **Tao Jin**†, **Wenlong Xu***, **Wang Lin**, **Yangyang Wu**

Zhejiang University
zrguo.cs@gmail.com

Abstract

Multimodal sentiment analysis (MSA) is an emerging research topic that aims to understand and recognize human sentiment or emotions through multiple modalities. However, in real-world dynamic scenarios, the distribution of target data is always changing and different from the source data used to train the model, which leads to performance degradation. Common adaptation methods usually need source data, which could pose privacy issues or storage overheads. Therefore, test-time adaptation (TTA) methods are introduced to improve the performance of the model at inference time. Existing TTA methods are always based on probabilistic models and unimodal learning, and thus can not be applied to MSA which is often considered as a multimodal regression task. In this paper, we propose two strategies: **C**ontrastive **A**daptation and **S**table **P**seudo-label generation (CASP) for test-time adaptation for multimodal sentiment analysis. The two strategies deal with the distribution shifts for MSA by enforcing consistency and minimizing empirical risk, respectively. Extensive experiments show that CASP brings significant and consistent improvements to the performance of the model across various distribution shift settings and with different backbones, demonstrating its effectiveness and versatility.

Code — https://github.com/zrguo/CASP

1 Introduction

Multimodal Sentiment Analysis (MSA) aims to understand and interpret human sentiment or emotions expressed through multiple modalities such as text, video and audio. Compared to traditional multimodal sentiment analysis which focuses on analyzing text data to determine the sentiment or emotion associated with a particular piece of text, MSA combines information from various modalities to gain a deeper understanding of sentiment. With the success of multimodal learning, MSA has attracted much attention (Zadeh et al. 2016; Tsai et al. 2019; Guo, Jin, and Zhao 2024). However, in real-world dynamic scenarios, the test data distribution is always changing which could lead to performance degradation of the model. For example, the model is trained on a multimodal sentiment analysis dataset in English but tested on a Chinese dataset, or the model is trained on a dataset with no background noise in the audio modality but tested on a dataset with a lot of background noise in the audio modality. Besides, in the video modality, different people have different facial traits. All of these things can be regarded as distribution shifts during the test stage and could lead to performance degradation of the model.

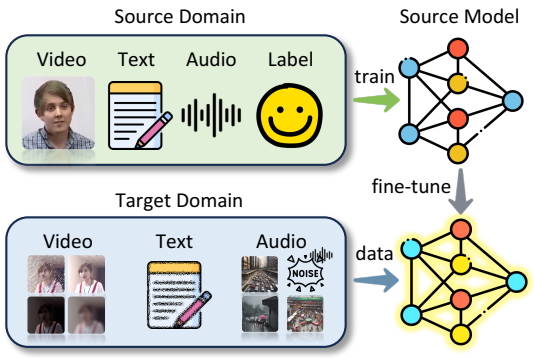

Figure 1: Test-time adaptation for multimodal sentiment analysis. The source domain data is used for source model training and is unavailable during the adaptation process. The target domain data is unlabeled.

To address distribution shifts at test time, Test-Time Adaptation (TTA) is proposed (Wang et al. 2020). TTA aims at overcoming the distribution gaps between source and target domains during test time without accessing the source data and the labels of the target data (Wang et al. 2020). Figure 1 presents the setting of TTA for MSA. However, existing methods can not be applied to MSA for two main reasons. On the one hand, MSA is often regarded as a regression task (Bagher Zadeh et al. 2018; Tsai et al. 2019) where the label is a score representing the intensity of the sentiment. However, most existing methods (Wang et al. 2020; Chen et al. 2022; Wang et al. 2022; Zhang, Levine, and Finn 2022) are based on probabilistic models for classification. For example, Wang et al. (2020) propose entropy minimization for adaptation which is a function for classification tasks and can not be applied to regression tasks. On the other hand, MSA is a multimodal task whereas existing methods mainly

*These authors contributed equally.
†Corresponding author

focus on unimodal tasks, overlooking the property of multimodal data and thus can not be applied to MSA. For instance, Zhang, Levine, and Finn (2022) perform different data augmentations on images to estimate marginal output distribution averaged over augmentations. However, it is hard to implement data augmentations on multimodal data, especially on extracted features instead of raw data.

Based on the above observations, in this paper, we propose **C**ontrastive **A**daptation and **S**table **P**seudo-label generation (CASP) for test-time adaptation for multimodal sentiment analysis. Specifically, our adaptation process has two stages: i) we introduce a contrastive adaptation strategy via modality random dropout to enforce consistency and improve the generalization ability of the model, meanwhile generating pseudo labels every few epochs and ii) we calculate the average value of the difference between the pseudo labels generated in stage one to select high-confident pseudo labels for self-training. The two stages deal with the TTA problem for MSA from two perspectives. Concretely, the contrastive adaptation strategy adapts the model by consistency regularization while the self-training with stable pseudo labels adapts the model by minimizing the empirical risk.

We conduct extensive experiments on three multimodal sentiment analysis datasets: CMU-MOSI (Zadeh et al. 2016), CMU-MOSEI (Bagher Zadeh et al. 2018) and CH-SIMS (Yu et al. 2020). We use different backbones to validate CASP's universality. The results show that CASP outperforms all the baselines significantly and consistently, demonstrating its superiority and versatility. Then, ablation experiments are conducted to measure the contribution of contrastive adaptation and stable pseudo labels and for a better understanding of CASP. To summarize, our contributions are as follows:

- We propose test-time adaptation techniques CASP for multimodal sentiment analysis to alleviate the distribution shifts between the source domain and target domain data. To the best of our knowledge, CASP is the first TTA method for *multimodal regression tasks*.
- We propose two novel strategies to address the distribution shifts of the target domain: contrastive adaptation to enforce consistency and stable pseudo-label generation to minimize the empirical risk.
- We show that CASP brings significant and consistent performance improvements to TTA for MSA across a range of settings and different backbones. The experimental results demonstrate the superiority and versatility of CASP.

2 Related Work

Multimodal Sentiment Analysis. Multimodal Sentiment Analysis (MSA) aims to predict sentiment intensity using multiple modalities such as text, video and audio. The main challenge of MSA is how to integrate information from different modalities effectively. Currently, there are mainly two types of fusion strategies: feature-level fusion (early fusion) and decision-level fusion (late fusion). Feature-level fusion methods (Lazaridou et al. 2015; Liang et al. 2018; Wang et al. 2019) combine the features extracted from different modalities to create a unified feature representation via concatenation or other methods before feeding it into the network. Different from feature-level methods, decision-level methods (Tsai et al. 2019; Yu et al. 2020) process different modalities separately and integrate them into the final decision. MISA (Hazarika, Zimmermann, and Poria 2020) projects each modality to two distinct subspaces to provide a holistic view of the multimodal data. MMIM (Han, Chen, and Poria 2021) hierarchically maximizes the mutual information in unimodal input pairs and between multimodal fusion result and unimodal input to maintain task-related information. UniMSE (Hu et al. 2022) proposes a knowledge-sharing framework that unifies MSA and MER to improve the performance. However, all these methods assume that the train and test data come from the same distribution. When there is a distribution shift between the train and test data, the performance of these models will degrade.

Test-time Adaptation. Test-Time Adaptation (TTA) refers to the adaptation of a pre-trained model to new target domain data without having access to the source domain data and the labels of target domain data. Unlike domain adaptation which requires access to both source and target data for adaptation, test-time adaptation methods do not require any data from the source domain and any label from the target domains. Among the various categories, one notable category is online TTA methods (Wang et al. 2020; Liang, Hu, and Feng 2020; Zhang, Levine, and Finn 2022). For example, TENT (Wang et al. 2020), the first TTA approach, takes a pre-trained model and adapts it to the test data by updating the trainable parameters in normalization layers using entropy minimization. Source Hypothesis Transfer (SHOT) (Liang, Hu, and Feng 2020) proposes to update only the encoder parameters and align source and target representation by entropy minimization and pseudo-labeling. Recently, some works have delved into the multimodal test-time adaptation (Shin et al. 2022; Yang et al. 2024). MM-TTA (Shin et al. 2022) proposes two complementary modules within and across the modalities to obtain reliable pseudo labels. READ (Yang et al. 2024) proposes reliable fusion against reliability bias and a novel objective function for robust multi-modal adaptation. In addition to online TTA methods, another category is robust TTA methods (Niu et al. 2022; Zhou et al. 2023). This kind of method takes some challenging issues into account such as single sample and label shifts. Recently, some researchers have started exploring continual TTA methods (Wang et al. 2022; Gan et al. 2023; Wang et al. 2024) which deal with the continually changing domain shifts in real-world scenarios.

However, the methods mentioned above all consider the probabilistic model for classification tasks. Therefore, they can not be applied to regression tasks such as sentiment analysis and image quality assessment. One recent work (Roy et al. 2023) proposes auxiliary tasks to enable TTA for regression tasks. However, it is a unimodal framework that needs an image augmentation strategy, which can not be applied to multimodal tasks. In contrast, this paper proposes a TTA method for multimodal regression tasks.

3 Methodology

In this section, we will introduce our proposed method CASP. First, we will formulate our problem setting in Section 3.1. Then, we will introduce a contrastive adaptation strategy

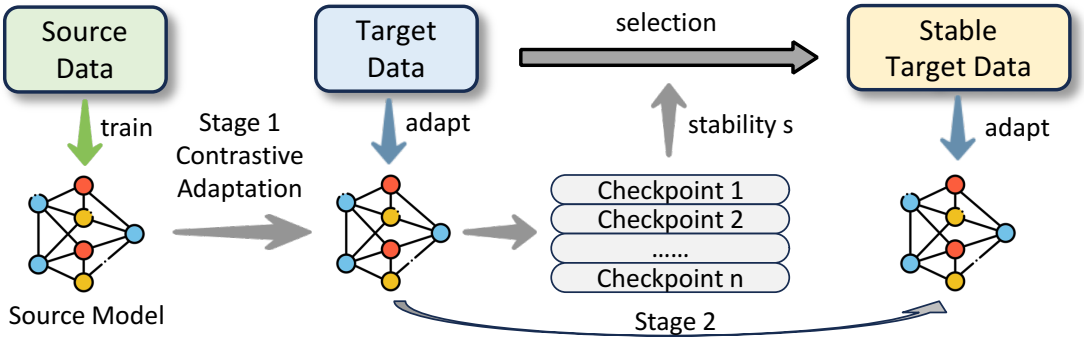

Figure 2: The overall framework of CASP. The adaptation process of CASP has two stages. Stage 1: contrastive adaptation to enforce consistency via modality random dropout. Stage 2: utilizing the checkpoints generated in Stage 1 to select high-confident pseudo labels for self-training. The two stages address the distribution shifts by consistency regularization and empirical risk minimization respectively.

via modality random dropout in Section 3.2. Finally, in Section 3.3, we will introduce a stable pseudo-label generation strategy. The overall framework of our method is presented in Figure 2.

3.1 Problem Formulation

In MSA tasks, there are usually three modalities: audio, video and text. Therefore, we define the source domain data as $\mathcal{S} = \{(\boldsymbol{s}_i, y_i)\}_{i=1}^{N_s}$ where N_s is the number of data and $\boldsymbol{s}_i = (s_i^a, s_i^v, s_i^t)$ represents the audio, video and text modality, respectively. In the TTA setting, we first pre-train our model on the source domain data $\mathcal{S}$. Suppose our model consists of the encoder $\mathcal{M}$ to get the feature representations and the prediction head $\mathcal{F}$ to get the final predictions, the output of the model is:

$$\hat{y} = \mathcal{F}_{\theta_f}(\mathcal{M}_{\theta_m}(\boldsymbol{s}_i)), \quad \boldsymbol{s}_i \in \mathcal{S} \tag{1}$$

where θ_m and θ_f are the parameters of $\mathcal{M}$ and $\mathcal{F}$, respectively. In MSA tasks, the loss function is often L1 loss (Tsai et al. 2019; Yu et al. 2020; Guo, Jin, and Zhao 2024). Therefore, the optimization process is:

$$\theta_m^*, \theta_f^* = \arg\min_{\theta_m, \theta_f} |\hat{y} - y| \tag{2}$$

Then we discard the source domain data and use the target domain data for adaptation. We define target domain data as $\mathcal{T} = \{\boldsymbol{x}_i\}_{i=1}^{N_t}$ where N_t is the number of data. The labels of the target domain are unavailable.

3.2 Contrastive Adaptation

During the adaptation process, the predictions of the model are expected to be consistent when different kinds of data augmentation strategies are implemented. Some existing TTA methods (Zhang, Levine, and Finn 2022; Wang et al. 2022) impose data augmentation strategies and calculate the average probability distribution of the augmented data to minimize the entropy or as pseudo labels. However, in multimodal regression tasks, we can neither calculate probability distributions nor perform data augmentation. Since many multimodal tasks use extracted features instead of raw data (Tsai et al. 2019; Yu et al. 2020; Guo, Jin, and Zhao 2024), data augmentation becomes difficult to implement. In order to enforce consistency, we introduce a contrastive adaptation strategy via modality random dropout to improve the generalization ability of the model. The overall diagram of the contrastive adaptation strategy is presented in Figure 3.

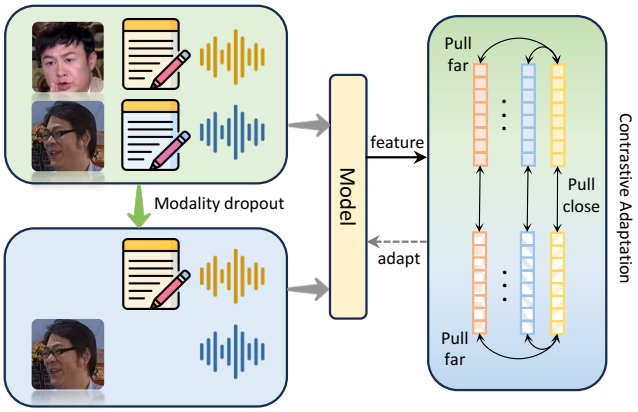

Figure 3: The overview of contrastive adaptation strategy. We randomly drop a modality to generate new data. Then we enforce the representations of the original data and the new data closer and distance the representation of the original data from the other representations in the batch.

Specifically, given $\boldsymbol{x}_i \in \mathcal{T}$, we impose a random dropout of modalities and replace the missing modality with $\mathbf{0}$ or other fixed value. We denote the data after modality random dropout as $\boldsymbol{x}_i^{\text{aug}}$, where different missing modality cases can be considered as different data augmentation strategies. We can obtain the feature representations $\boldsymbol{h}$ of $\boldsymbol{x}_i$ and $\boldsymbol{x}_i^{\text{aug}}$ following:

$$\boldsymbol{h}_i = \mathcal{M}(\boldsymbol{x}_i), \quad \boldsymbol{h}_i^{\text{aug}} = \mathcal{M}(\boldsymbol{x}_i^{\text{aug}}) \tag{3}$$

To impose consistency regularization, we want to bring $\boldsymbol{h}_i$ and $\boldsymbol{h}_i^{\text{aug}}$ closer together and move $\boldsymbol{h}_i$ and other representations in the batch further apart. Therefore, we consider a

modified *NT-Xent* loss (Chen et al. 2020). Let $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v}/||\boldsymbol{u}||||\boldsymbol{v}||$ denote the dot product between ℓ_2 normalized $\boldsymbol{u}$ and $\boldsymbol{v}$ (*i.e.* cosine similarity). The loss function for a positive example $(\boldsymbol{h}_i, \boldsymbol{h}_i^{\text{aug}})$ is as follows:

$$\ell_{\boldsymbol{h}_i, \boldsymbol{h}_i^{\text{aug}}} = -\log \frac{\exp(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_i^{\text{aug}})/\tau)}{\sum_{k=1}^{K} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{h}_i, \boldsymbol{h}_k)/\tau)} \quad (4)$$

where K is the batch size, τ is a temperature parameter, and $\mathbb{1}_{[k \neq i]} \in \{0, 1\}$ is a sign function evaluating to 1 if $k \neq i$. The total loss can be written as:

$$\mathcal{L} = \frac{1}{2K} \sum_{k=1}^{K} (\ell_{\boldsymbol{h}_k, \boldsymbol{h}_k^{\text{aug}}} + \ell_{\boldsymbol{h}_k^{\text{aug}}, \boldsymbol{h}_k}) \quad (5)$$

For stability and efficiency, previous works (Wang et al. 2020; Roy et al. 2023) usually reconcile the distribution shifts by updating normalization layers. In our method, we follow these works, only updating the linear and lower-dimensional feature modulation parameters (*i.e.* normalization layers). Through contrastive adaptation strategy, the model will learn more generalizable features and become more consistent.

3.3 Stable Pseudo-label Generation

Some previous TTA methods (Wang et al. 2022; Zhang, Levine, and Finn 2022) propose to generate high-confident pseudo labels for entropy minimization. Due to the probabilistic models, it is easy to generate pseudo labels to measure confidence (*i.e.* the probability of each class). However, the output of a regression model is non-probabilistic, thus making it hard to measure confidence. In this subsection, we propose to measure the confidence of pseudo labels via a dynamic training process. Concretely, in Section 3.2, we introduce a contrastive adaptation strategy to improve the generalization ability of the model and make the model more consistent. Suppose the model is adapted for E epochs, we denote the model of epoch e as $F_e = (\mathcal{M}_{\theta_m^e}, \mathcal{F}_{\theta_f^e})$ where $e = 1, 2, \cdots, E$. Then the predictions of the model F_e can be expressed as:

$$\hat{y}_e = F_e(\boldsymbol{x}) \quad (6)$$

where $\hat{y}_e$ represents the predictions of all the training samples using F_e. For every epoch e, we can calculate the pseudo labels $\hat{y}_e$. During this dynamic process, some of the pseudo labels change a lot while some of them remain stable. Based on this observation, we propose to select those stable pseudo labels as high-confident pseudo labels and those that change a lot as low-confident pseudo labels. For efficiency and high quality, we calculate the pseudo labels of all the training samples every M epochs. Mathematically, we calculate the difference between every two consecutive checkpoints, and use the average difference to measure the stability s:

$$s = \frac{1}{\lfloor \frac{E}{M} \rfloor} \sum_{i=0}^{\lfloor \frac{E}{M} \rfloor - 1} |\hat{y}_{iM} - \hat{y}_{(i+1)M}| \quad (7)$$

where $\lfloor \frac{E}{M} \rfloor$ represents the largest integer not exceeding E/M and $\hat{y}_0$ denotes the labels generated by the source model. After obtaining the stability s, we set a threshold to select high-confident pseudo labels. Specifically, we use λ-quantiles as the threshold. When s is smaller than the threshold, we select the sample as our self-training sample. When s is larger than the threshold, we discard the sample. For the values of pseudo labels, we use the average value of $\hat{y}$ of all the checkpoints:

$$\tilde{y} = \frac{1}{\lfloor \frac{E}{M} \rfloor + 1} \sum_{i=0}^{\lfloor \frac{E}{M} \rfloor} \hat{y}_{iM} \quad (8)$$

Then, we can obtain the self-training dataset $\mathcal{T}_{\text{train}} = \{\boldsymbol{x}_i, \tilde{y}_i\}_{i=1}^{N_{\text{train}}}$ where N_{train} is the number of selected high-confident samples. Then we use $\mathcal{T}_{\text{train}}$ for training.

4 Experiments

4.1 Datasets and Evaluation Metrics

CMU-MOSI (Zadeh et al. 2016) is a popular dataset for multimodal (audio, text and video) sentiment analysis. It comprises 93 English YouTube videos, containing 89 distinct speakers, including 41 female and 48 male speakers. Each segment is manually annotated with a sentiment score ranging from strongly negative to strongly positive (-3 to +3).

CMU-MOSEI (Bagher Zadeh et al. 2018) is an extension of CMU-MOSI. It contains more than 65 hours of annotated video from more than 1000 speakers and 250 topics. It has a total number of 3,228 videos which are divided into 23,453 segments. Compared with CMU-MOSI, it covers a wider range of topics.

CH-SIMS (Yu et al. 2020) is a Chinese multimodal sentiment analysis dataset that has three modalities (audio, text and video). It has a total of 60 videos which contains 2,281 refined segments in the wild annotated with a sentiment score ranging from strongly negative to strongly positive (-1 to 1).

For all three datasets, we use binary accuracy (ACC), F1 score (F1) and mean absolute error (MAE) as evaluation metrics.

4.2 Baselines

To the best of our knowledge, we are the first to introduce the TTA method to multimodal regression tasks. Therefore, previous methods can not be applied to MSA tasks due to the properties of multimodal data and the non-probabilistic model. In our experiments, we mainly compare our method with five baselines: **Source** is the model pre-trained on the source domain data. Then it is tested on the target domain data without any adaptation strategy. **ST** is the self-training method. We use the source model to generate pseudo labels of the target domain data. Then we use these pseudo labels to train the source model. **Norm** is also the self-training method. Different from ST where we train the whole model, Norm only trains the normalization layers and freezes other parameters. This method (Wang et al. 2020; Roy et al. 2023) is commonly used in TTA. **GC** is the group contrastive strategy proposed in TTA-IQA (Roy et al. 2023). In TTA-IAQ, the authors propose a group contrastive strategy and rank loss strategy. However, the rank loss strategy can not be applied to multimodal data due to the image augmentation strategy. Therefore, we only apply GC to our task. **RF** is the reliable

Backbone	Method	MOSEI→SIMS			MOSI→SIMS			MOSI→MOSEI			SIMS→MOSI			SIMS→MOSEI		
		ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE	ACC	F1	MAE
Late Fusion	Source	60.96	63.09	2.01	39.17	39.12	2.10	66.57	67.42	1.25	40.12	45.46	2.18	47.14	57.47	1.77
	ST	62.01	65.19	1.95	40.48	39.55	2.05	67.41	67.90	1.23	40.41	46.35	**2.00**	47.34	58.35	1.87†
	Norm	61.40	64.38	2.04†	38.51†	38.84†	2.12†	66.62	67.53	1.30†	40.27	47.22	2.30†	47.70	58.74	1.84†
	GC (Roy et al. 2023)	62.62	65.38	1.98	42.23	42.89	1.97	67.03	67.83	1.25	40.94	46.87	2.21†	47.45	59.07	1.76
	RF (Yang et al. 2024)	61.12	64.07	1.97	40.19	40.01	2.06	67.11	67.70	1.28†	40.18	45.98	2.28†	47.61	58.44	1.86†
	CASP	**64.23**	**67.75**	**1.81**	**51.27**	**53.15**	**1.73**	**69.12**	**69.17**	**0.96**	**48.03**	**50.43**	2.04	**49.09**	**59.11**	**1.60**
Early Fusion	Source	45.95	45.28	2.15	36.76	37.83	2.42	66.75	67.35	1.24	40.17	40.60	1.75	46.39	50.61	1.34
	ST	48.80	47.20	2.11	34.79†	36.52†	2.50†	66.63†	67.35	1.34†	41.74	42.39	1.55	47.14	53.68	1.32
	Norm	43.76†	44.06†	2.25†	36.23†	37.94	2.47†	66.98	67.54	1.30†	43.95	43.40	1.56	45.80†	48.13†	1.29
	GC (Roy et al. 2023)	47.64	47.60	2.10	37.22	37.70†	2.29	67.12	67.40	1.22	42.67	43.08	1.54	46.77	49.12†	1.30
	RF (Yang et al. 2024)	46.12	46.25	2.18†	35.18†	35.97†	2.46†	66.84	67.39	1.27†	42.58	43.02	1.59	46.61	50.39†	1.31
	CASP	**63.89**	**66.43**	**1.80**	**40.12**	**41.65**	**2.06**	**68.32**	**68.90**	**1.08**	**46.57**	**47.10**	**1.44**	**47.90**	**57.13**	**1.26**

Table 1: Quantitative results across five different distribution shift settings with two different backbones. For simplicity, we use MOSI, MOSEI and SIMS to represent CMU-MOSI, CMU-MOSEI and CH-SIMS. **Bold**: best results. Underline: second best results. † represents that the performance decreases compared with the source model without adaptation. We report the average results using five different random seeds.

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
Epoch 0	-2.38	+0.83	+2.12	-1.21	+1.98	+1.68
Epoch 3	-2.39	+0.32	+1.17	-1.21	+1.99	+1.74
Epoch 6	-2.40	-0.17	+0.85	-1.22	+1.95	+1.20
Epoch 9	-2.41	+0.08	+1.25	-1.23	+2.11	+1.27
Epoch 12	-2.41	+1.02	-0.01	-1.23	+2.12	+1.01
Epoch 15	-2.42	+1.14	+0.18	-1.23	+2.11	+0.84
Stability s	0.008	0.46	0.62	0.004	0.05	0.22
GT	-3.0	+2.4	-0.8	-1.8	+3.0	-0.8
Selected?	✔	✗	✗	✔	✗	✗
Pseudo Label	-2.40	-	-	-1.22	-	-

Table 2: Case study of pseudo-label generation process on MOSEI→SIMS. We adapt the model for 15 epochs and set interval parameter $M = 3$. The table shows the predictions of six samples. "GT" denotes ground truth. "Selected?" denotes whether the sample is selected for self-training. The threshold is 0.012 as shown in Figure 4 when $\lambda = 95$.

fusion strategy proposed in READ (Yang et al. 2024). READ proposes two strategies: reliable fusion and robust adaptation. Reliable fusion strategy is a new paradigm that modulates the attention between modalities in a self-adaptive way. Robust adaptation is based on probabilistic models which can not be applied to regression tasks. Therefore, we use the reliable fusion strategy for comparison.

4.3 Implementation Details

Raw Feature Extraction. For text modality, we use pre-trained BERT (Devlin et al. 2019) to obtain word embeddings. We use BERT-base for CMU-MOSI and CMU-MOSEI and Chinese BERT-base for CH-SIMS. Each word is represented as a 768-dimensional vector. For audio modality, we use LibROSA (McFee et al. 2015) to extract features. For video modality, we extract face features using OpenFace 2.0 (Baltrusaitis et al. 2018) toolkit.

Source Domain and Target Domain. We denote source domain to target domain as $A \rightarrow B$ where A represents the source domain and B is the target domain. We validate CASP across five distribution shift settings: CMU-MOSEI→CH-SIMS, CMU-MOSI→CH-SIMS, CMU-MOSI→CMU-MOSEI, CH-SIMS→CMU-MOSI and CH-SIMS→CMU-MOSEI. We do not use CMU-MOSEI→CMU-MOSI because CMU-MOSEI is an extension of CMU-MOSI and covers a wider range of topics compared with CMU-MOSI.

Backbones. To demonstrate the generalization ability of our method, we use two different backbones: the feature-level fusion (early fusion) method and the decision-level fusion (late fusion) method. Specifically, we use the transformer encoder (Vaswani et al. 2017) as the backbone. For fairness, we use the same backbone for all the methods.

Training Details. For source domain pre-training and contrastive adaptation, we use the AdamW optimizer with a learning rate of $1e-3$. We adapt the model for 15 epochs and the interval hyperparameter M is set to 3. For stable pseudo-label generation, we set the threshold hyperparameter λ as 95. For self-training using stable pseudo labels, we use the AdamW optimizer with a learning rate of $5e-4$ and train the model for 5 epochs. The batch size of all the experiments is 24. Besides, we use gradient clipping and set the threshold as 0.8. We also use a step scheduler with a step size of 10 and decay rate $\gamma = 0.1$. To avoid randomness, we train the model five times using five different random seeds and report the average results.

4.4 Main Results

We present our quantitative results across five different distribution shift settings with two different backbones in Table 1. Compared with the source model without any adaptation strategy, CASP brings significant and consistent performance improvements across all settings and with different backbones. ST, Norm, GC and RF are four different methods to mitigate the distribution shifts between the source domain and the target domain. However, we observe that all these four methods have performance degradation on some metrics across some distribution shifts, which are marked with † in the table. Only CASP can bring consistent performance improvements.

Besides, the source model performs very poorly

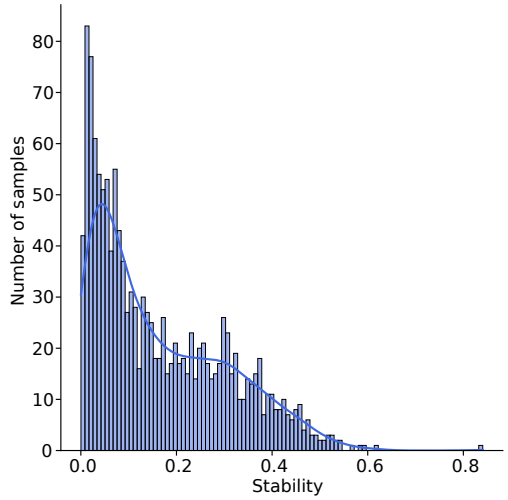

Figure 4: The distribution of stability s on MOSEI→SIMS.

on MOSEI→SIMS, MOSI→SIMS, SIMS→MOSI and SIMS→MOSEI because SIMS is a Chinese multimodal sentiment analysis dataset while MOSI and MOSEI are English multimodal datasets. Therefore, SIMS has a huge distribution shift from MOSI and MOSEI. The accuracies across these settings are below 50%. All the baselines bring limited improvements across these settings while CASP brings significant improvements. On MOSEI→SIMS, CASP improves the accuracy of the early fusion backbone by nearly 20% while the second best method ST improves the accuracy by just around 3%. On MOSI→SIMS, CASP improves the accuracy of the late fusion backbone by around 11% while the second best method GC improves the accuracy by just around 3%. Only on SIMS→MOSI, ST performs better on MAE with 0.04 higher than that of CASP. These results fully demonstrate the superiority and versatility of CASP.

4.5 Stable Pseudo-label Generation

In Figure 4, we present the distribution of stability s on MOSEI→SIMS. As shown in the figure, most of the samples do not change drastically during the contrastive adaptation process. When we use the most stable 5% of the samples as pseudo labels (i.e. the hyperparameter $\lambda = 95$), the stability threshold is 0.012. We present the changes of six samples during the contrastive adaptation process in Table 2. From Case 2, Case 3 and Case 6, we can find that these three cases have high stability s (*i.e.* not stable) and their predictions differ greatly from the ground truths. This demonstrates that high stability means low-confident labels, indicating the effectiveness of our method in choosing high-confident pseudo labels. Comparing Case 1 and Case 5, we can find that although Case 5 also has low stability s, the difference between the prediction and the ground truth of Case 5 is bigger than that of Case 1, indicating that lower stability means higher confidence. Moreover, from epoch 0 (*i.e.* Source model) to epoch 15, it is easy to find that the predictions change towards the ground truths. For example, at epoch 0, the prediction of Case 1 is -2.38, and at epoch 15, the prediction is -2.42, which is closer to the ground truth -3.0. This also demonstrates the effectiveness of our contrastive adaptation.

Backbone	CA	SPL	T_1	T_2	T_3	T_4	T_5
Late Fusion	✗	✗	60.96	39.17	66.57	40.12	47.14
	✓	✗	62.36	49.45	67.74	41.15	48.33
	✓	✓	64.23	51.27	69.12	48.03	49.09
Early Fusion	✗	✗	45.95	36.76	66.75	40.17	46.39
	✓	✗	61.62	38.95	67.76	44.10	47.23
	✓	✓	63.89	40.12	68.32	46.57	47.90

Table 3: Quantitative results of contributions of the contrastive adaptation (CA) and stable pseudo-label generation (SPL). T_1, T_2, T_3, T_4, T_5 represents the distribution shift: MOSEI→SIMS, MOSI→SIMS, MOSI→MOSEI, SIMS→MOSI and SIMS→MOSEI, respectively. We only report the accuracy in the table.

Backbone	n	T_1	T_2	T_3	T_4	T_5
Late Fusion	none	60.96	39.17	66.57	40.12	47.14
	1	62.36	49.45	67.74	41.15	48.33
	2	61.11	44.67	66.94	40.29	47.60
Early Fusion	none	45.95	36.76	66.75	40.17	46.39
	1	61.62	38.95	67.76	44.10	47.23
	2	53.12	36.94	66.81	41.57	46.80

Table 4: The impact of the number of dropped modalities. The table shows the accuracy of the model. n represents the number of dropped modalities. T_1, T_2, T_3, T_4 and T_5 have the same meaning as Table 3.

4.6 Ablation Study

In this subsection, we conduct a series of ablation experiments for a better understanding of CASP. Concretely, we will analyze the contribution of contrastive adaptation and stable pseudo-label generation. Besides, we will explore the impact of the number of dropped modalities, the quality of the selected pseudo labels, the impact of the interval hyperparameter M and the stability threshold λ on the performance of the model.

Contributions of contrastive adaptation and stable pseudo-label generation. Table 3 presents the accuracy of the ablation experiments. We observe that both the contrastive adaptation strategy and stable pseudo-label generation improve the performance of the model. Particularly, compared to the stable pseudo-label generation strategy, the contrastive adaptation strategy has more potential to improve the performance of the model. Across five different distribution shifts, self-training with stable pseudo labels improves the accuracy by around 1%-7% while the contrastive adaptation strategy can improve the accuracy by up to around 10%-15%. Furthermore, we observe that contrastive adaptation helps more when the accuracy of the source model is low and helps less when the accuracy of the source model is high. In contrast, the stable pseudo-label strategy brings relatively steady and consistent improvements to the performance of the model.

Backbone	Method	T_1	T_2	T_3	T_4	T_5
Late Fusion	SPL ($M=1$)	40.25	37.25	47.08	44.98	52.01
	SPL ($M=2$)	41.07	38.55	47.60	45.34	55.17
	SPL ($M=3$)	42.11	38.99	48.86	45.22	57.81
	SPL ($M=4$)	42.23	38.89	49.55	46.80	59.21
Early Fusion	SPL ($M=1$)	46.11	45.20	46.14	52.01	51.12
	SPL ($M=2$)	46.85	45.56	46.55	52.17	51.02
	SPL ($M=3$)	46.61	46.06	47.69	54.70	52.48
	SPL ($M=4$)	46.80	46.23	47.71	54.57	53.96

Table 5: The impact of the interval hyperparameter M on the quality of the stable pseudo labels. The table reports the rates of MAE decline compared to the pseudo labels obtained directly using the source model. In all the experiments, we fix the adaptation epoch $E = 20$. T_1, T_2, T_3, T_4 and T_5 have the same meaning as Table 3.

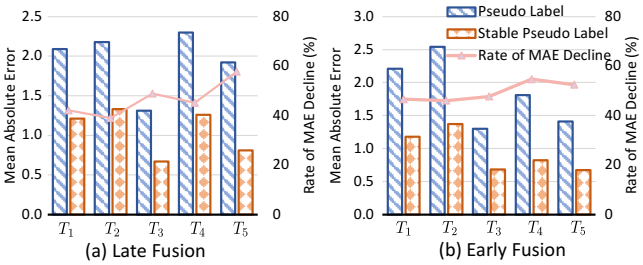

Figure 5: Effectiveness of the stable pseudo-label generation strategy across five different distribution shift settings and with two different backbones. T_1, T_2, T_3, T_4 and T_5 have the same meaning as Table 3.

The number of dropped modalities. In the contrastive adaptation strategy, we randomly drop a modality as a way of augmentation. To explore the impact of the number of dropped modalities, we conduct experiments and present the results in Table 4. Both $n = 1$ and $n = 2$ bring consistent improvements to the performance of the model. Furthermore, we observe that the performance of the model only dropping one modality is better than that of the model dropping two modalities. In our datasets, there are only three modalities. Therefore, dropping two modalities may lose a lot of information and enforce strong consistency, which may hinder the performance of the model compared with only dropping one modality. If a dataset has more modalities such as five modalities, dropping two modalities may be a better choice.

Quality of the stable pseudo labels. To demonstrate that the quality of our stable pseudo labels is much better than that of pseudo labels obtained directly using the source model, we calculate the mean absolute error between the pseudo labels and the ground truths and present our results in Figure 5. From the figure, we observe that the MAE of stable pseudo labels is much lower than that of pseudo labels obtained using the source model. The rates of MAE decline across five different distribution shift settings and two different backbones are all around 40%-60%, demonstrating the effectiveness of our stable pseudo-label generation strategy.

The impact of the interval hyperparameter M. To explore the impact of the interval hyperparameter M on the quality of stable pseudo labels, we select $M = 1, 2, 3, 4$ and generate stable pseudo labels to calculate the rates of MAE decline compared to the pseudo labels generated by the source model. We report our results in Table 5. We observe that the quality of the stable pseudo labels increases as M increases. Intuitively, when the interval M is large, the difference between the two checkpoints is large. This is beneficial to the selection of high-confident labels because our stable pseudo-label generation strategy calculates the average value of the difference between two consecutive checkpoints and selects the labels whose values are lower than a threshold. A large M means large differences and thus would help to select these stable labels. However, to include relatively more checkpoints, a large M requires more training epochs E, which could increase the time of the adaptation process. To get a balance between the adaptation time and the performance, $M = 2, 3$ would be an opportune value.

Backbone	λ	T_1	T_2	T_3	T_4	T_5
Late Fusion	50	12.01	10.17	13.22	13.67	19.90
	75	30.66	21.85	27.68	31.02	41.10
	95	42.11	38.99	48.86	45.22	57.81
Early Fusion	50	13.17	12.01	13.22	19.34	17.25
	75	26.12	22.20	23.41	31.46	30.14
	95	46.61	46.06	47.69	54.70	52.48

Table 6: The impact of the threshold λ on the quality of the stable pseudo labels. The table reports the rates of MAE decline compared to the pseudo labels generated using the source model. T_1, T_2, T_3, T_4 and T_5 have the same meaning as Table 3.

The impact of the threshold λ. We select three different threshold λ and present our ablation results in Table 6. The results demonstrate our stable pseudo-label generation strategy can generate high-confident pseudo labels whatever λ is. With the increase of λ, the quality of the pseudo labels also increases. However, with the increase of λ, the number of samples for self-training decreases. Therefore, if the test set has many samples, we are expected to choose a large λ while if the test set does not have many samples, we can decrease the value of λ.

5 Conclusion

In this paper, we focus on the test-time adaptation for multimodal sentiment analysis. Due to the reason that multimodal sentiment analysis is a multimodal regression task, existing methods can not be applied. To address the distribution shifts for multimodal sentiment analysis, we propose contrastive adaptation and stable pseudo-label generation (CASP) strategies. CASP has two stages: contrastive adaptation to enforce consistency and self-training with stable pseudo labels to minimize empirical risk. We conduct extensive experiments across various distribution shift settings and with different backbones. The results demonstrate the superiority and versatility of CASP. Ablation experiments are then conducted to validate the main components of CASP.

References

Bagher Zadeh, A.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246. Melbourne, Australia: Association for Computational Linguistics.

Baltrusaitis, T.; Zadeh, A.; Lim, Y. C.; and Morency, L.-P. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, 59–66.

Chen, D.; Wang, D.; Darrell, T.; and Ebrahimi, S. 2022. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 295–305.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 1597–1607. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Gan, Y.; Bai, Y.; Lou, Y.; Ma, X.; Zhang, R.; Shi, N.; and Luo, L. 2023. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7595–7603.

Guo, Z.; Jin, T.; and Zhao, Z. 2024. Multimodal Prompt Learning with Missing Modalities for Sentiment Analysis and Emotion Recognition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1726–1736.

Han, W.; Chen, H.; and Poria, S. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9180–9192.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 1122–1131. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Hu, G.; Lin, T.-E.; Zhao, Y.; Lu, G.; Wu, Y.; and Li, Y. 2022. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7837–7851. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lazaridou, A.; et al. 2015. Combining Language and Vision with a Multimodal Skip-gram Model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Liang, J.; Hu, D.; and Feng, J. 2020. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, 6028–6039. PMLR.

Liang, P. P.; Liu, Z.; Zadeh, A. B.; and Morency, L.-P. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 150–161.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P. W.; McVicar, M.; Battenberg, E.; and Nieto, O. 2015. librosa: Audio and Music Signal Analysis in Python. In *SciPy*.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2022. Towards Stable Test-time Adaptation in Dynamic Wild World. In *The Eleventh International Conference on Learning Representations*.

Roy, S.; Mitra, S.; Biswas, S.; and Soundararajan, R. 2023. Test Time Adaptation for Blind Image Quality Assessment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16742–16751.

Shin, I.; Tsai, Y.-H.; Zhuang, B.; Schulter, S.; Liu, B.; Garg, S.; Kweon, I. S.; and Yoon, K.-J. 2022. Mm-tta: Multimodal test-time adaptation for 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16928–16937.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569. Florence, Italy: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2020. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *International Conference on Learning Representations*.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.

Wang, Y.; Hong, J.; Cheraghian, A.; Rahman, S.; Ahmedt-Aristizabal, D.; Petersson, L.; and Harandi, M. 2024. Continual Test-time Domain Adaptation via Dynamic Sample Selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1701–1710.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7216–7223.

Yang, M.; Li, Y.; Zhang, C.; Hu, P.; and Peng, X. 2024. Test-time adaption against multi-modal reliability bias. In *The Twelfth International Conference on Learning Representations*.

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 Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *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, M.; Levine, S.; and Finn, C. 2022. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35: 38629–38642.

Zhou, Z.; Guo, L.-Z.; Jia, L.-H.; Zhang, D.; and Li, Y.-F. 2023. ODS: test-time adaptation in the presence of open-world data shift. In *International Conference on Machine Learning*, 42574–42588. PMLR.