*Journal of Artificial Intelligence and Technology,* (Ahead of Print)
https://doi.org/10.37965/jait.2026.1022

ISTP
RESEARCH ARTICLE

# Test-Time Adaptation and Domain Generalization in Multimodal Sentiment Analysis: From Laboratory to Real-World Deployment

**Natalie Jie Yin Tan, Kathleen Swee Neo Tan, and Chi Wee Tan**
Tunku Abdul Rahman University of Management and Technology, Kuala Lumpur, Malaysia

(Received 19 November 2025; Revised 21 September 2026; Accepted 01 October 2026; Published online 02 November 2026)

***Abstract:*** The deployment of multimodal sentiment analysis (MSA) systems in real-world environments is hindered by domain shift, where performance degrades significantly when test data diverges from the training distribution. This review paper provides a comprehensive survey of two pivotal strategies to address this challenge: domain generalization (DG) and test-time adaptation (TTA). We present a novel taxonomy of state-of-the-art methods, categorizing them into data manipulation, representation learning, meta-learning, and hybrid approaches. The review critically examines techniques specifically designed for multimodal data, including modality-specific adaptation and cross-modal consistency enforcement. Furthermore, we discuss the common benchmarks and evaluation metrics that are used to assess robustness to real-world challenges such as missing modalities, not independent and identically distributed (non-i.i.d.) data streams, and computational constraints. Finally, the paper outlines critical open challenges and future research directions, emphasizing the need to handle mixed and continual shifts, adapt large multimodal models efficiently, and improve explainability. This work assists researchers in building reliable MSA systems for real-world deployment.

***Keywords:*** domain generalization; multimodal sentiment analysis; test-time adaptation

## I. INTRODUCTION

### A. EMERGENCE OF MULTIMODAL SENTIMENT ANALYSIS

Multimodal sentiment analysis (MSA) extends sentiment interpretation beyond text to include audio and visual cues, capturing richer emotional signals like sarcasm, tone, or body language that unimodal approaches are unable to detect [1]. For instance, the phrase "*It is OK*" may convey negativity if spoken with a strained tone or a forced smile, which text alone would fail to capture [2].

MSA addresses the shortcomings of text-only analysis, particularly in social media and real-world contexts where non-verbal signals are abundant. Integrating complementary cues from text, speech, and facial expressions enhances prediction accuracy and robustness [3], establishing MSA as a promising approach for reliable emotion recognition. The key challenge lies in designing effective fusion strategies that can integrate heterogeneous data while accounting for differing information densities across modalities [4].

### B. DOMAIN SHIFT IN MSA

A major limitation in current MSA systems is their reliance on the assumption that training and test data share the same distribution [5]. In practice, models often face *domain shift*, where input distributions at deployment diverge significantly from the training data. This results in a marked decline in model performance, particularly when models trained on clean benchmarks (e.g., CMU-MOSI and MOSEI) are applied to noisy, dynamic real-world streams [6].

Domain shift in MSA is complex because there may be different combinations of modalities present and expressive patterns vary across domains, as detailed in Section IV.A. Moreover, real-world deployments often encounter continually evolving target domains, leading to catastrophic forgetting or instability when models adapt without supervision [7]. These challenges point to the need for robust domain generalization (DG) and adaptation strategies for MSA.

### C. THE SOLUTION SPACE

Two main paradigms address domain shift: **DG** and **test-time adaptation (TTA)**. DG trains models on one or more source domains to perform robustly on unseen target domains by learning domain-invariant features. Variants such as single-source DG (SSDG) push this further by requiring generalization from a single domain [8], though universal models may not always be optimal for all test distributions.

TTA, by contrast, adapts a pretrained model at inference without access to source data or target labels. This is particularly relevant in privacy-sensitive or resource-limited scenarios [9]. TTA variants include **(i) fully offline TTA**, which adapts before inference [5]; **(ii) online TTA**, which adapts while predicting [10]; and **(iii) continual TTA (CTTA)**, where the model faces a stream of shifting domains [7]. Common techniques include pseudo-labeling, entropy minimization, consistency regularization, and normalization layer adjustment [11]. Despite promising results, TTA remains challenged by noisy supervision, catastrophic forgetting, and unstable convergence under not independent and identically distributed (non-i.i.d.) test streams [12]. Table I shows the differences between DG and TTA from various aspects [13].

### D. MOTIVATION AND CONTRIBUTION

This review seeks to bridge the gap between controlled laboratory research and real-world deployment of MSA. By surveying DG

Corresponding author: Natalie Jie Yin Tan (e-mail: natalietjy-wm22@student.tarc.edu.my).

© The Author(s) 2026. This is an open access article published under the CC BY license (https://creativecommons.org/licenses/by/4.0/). 

and TTA, we provide a structured analysis of how these methods can enhance MSA reliability in practice. Our contributions are as follows:

- **First dedicated review of robustness in MSA**: To the best of our knowledge, this is the first survey that focuses specifically on TTA and DG in MSA. Existing reviews primarily address fusion architectures or applications, leaving robustness underexplored.
- **A unified taxonomy of methods**: We introduce a taxonomy that organizes robustness approaches into DG, TTA, hybrid, and multimodal-specific strategies, further distinguishing key subfamilies such as entropy minimization, cross-modal consistency, and modality-specific adaptation.
- **Critical dataset and evaluation analysis**: We examine widely used MSA benchmarks and evaluation protocols, highlighting over-reliance on clean datasets and recommending robustness-focused benchmarks (e.g., MOSI-ASR, MVSA↔Yelp) to better reflect real-world deployment.
- **Identification of open challenges**: We discuss unresolved issues such as adaptation for regression tasks, robustness to modality corruption, handling continuous domain shifts, and the broader translation gap between laboratory-trained models and practical applications.

By synthesizing the current literature, this review provides both a foundation and a roadmap for building robust multimodal sentiment systems capable of real-world deployment.

The rest of the paper is organized as follows. Section II reviews the related work and the core concepts. Subsequently, Section III presents a taxonomy of DG, TTA, hybrid, and multimodal-specific methods. Sections IV and V examine the real-world challenges and deployment issues, as well as the datasets and evaluation. Lastly, the paper discusses other open challenges and future directions in Section VI.

# II. BACKGROUND AND CORE CONCEPTS

## A. MSA ARCHITECTURE'S SUSCEPTIBILITY TO DOMAIN SHIFT

As established in Section I.A, the standard MSA pipeline is effective—comprising feature extraction, fusion, and prediction—but remains inherently vulnerable to domain shift, limiting its real-world applicability. A general overview of this pipeline is shown in Fig. 1. The following sections examine feature extraction shifts, the sensitivity of fusion mechanisms, the prediction heads, and the limitations of evaluation metrics on clean benchmarks.

***1). FEATURE EXTRACTION AS A SOURCE OF DISTRIBUTIONAL SHIFT.*** Feature extractors are the primary vectors for domain shift. Models pretrained on generic corpora (e.g., BERT [14] for text, wav2vec [15] for audio, ResNet [16] for visuals) learn source-specific representations. Consequently, their performance degrades when faced with target domain variations such as new accents, dialects, lighting conditions, or cultural expressions [9]. This creates a distributional mismatch at the input level to the fusion module.

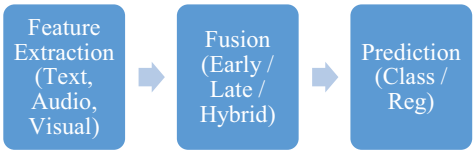



**Fig. 1.** The standard MSA pipeline. Its performance degrades under domain shift due to vulnerabilities at each stage.

***2). FUSION STRATEGIES AND THEIR SENSITIVITY TO CHANGING MODALITY DYNAMICS.*** Fusion mechanisms integrate features to capture cross-modal interactions [17]. However, they are typically optimized for the correlation patterns present in the source domain. These patterns are fragile; a domain shift—such as the introduction of ASR errors corrupting the text modality or background noise affecting audio—can sever these learned relationships, leading to incorrect fusion and performance failure. Table II summarizes common fusion categories and their inherent robustness considerations [18,19].

***3). PREDICTION HEADS AND EVALUATION.*** The final prediction head (classification or regression) operates on the fused feature vector. A shift in the distribution of this vector—caused by upstream extraction or fusion—directly causes miscalibration and inaccurate sentiment predictions [20,21]. Consequently, standard performance metrics (e.g., accuracy, F1-score, mean absolute error (MAE)) that are measured on clean, i.i.d. benchmarks fail to capture a model's resilience to the domain shifts prevalent in real-world deployment.

In summary, the conventional MSA architecture is sensitive to changes in modality dynamics and domain-specific distributions. Its assumption of stable feature distributions and static inter-modal relationships is routinely violated in practice, necessitating DG and TTA strategies explored in the remainder of this review.

## B. DOMAIN GENERALIZATION

DG aims to overcome a fundamental limitation of deep learning models—the assumption that training and test data are identically distributed—by learning models that can generalize effectively to entirely unseen target domains. This is critical for real-world deployment, where such domain shifts are the norm, rather than the exception. The paradigm has two constraints: models cannot access target data during training and cannot access source data during testing [7]. DG is typically studied in two settings: the more challenging SSDG, with only one source domain for training, and multi-source DG, which leverages multiple heterogeneous source domains [10]. A primary goal is to learn domain-invariant feature

**Table I.** Differences between DG and TTA

| Aspect | DG | TTA |
|---|---|---|
| **Timing** | During training. | At inference. |
| **Data** | Uses labeled source domains. | Uses unlabeled target data. |
| **Goal** | Universal model for broad unseen domains. | Specialized adaptation to current target data. |
| **Mechanism** | Learns domain-invariant features or meta-learning strategies. | Updates model parameters using objectives like pseudo-labeling or entropy minimization. |

**Table II.** Common fusion methods and their robustness profiles

| Category | Description | Robustness consideration |
|---|---|---|
| **Early fusion** | Concatenate raw or low-level features from different modalities into a single vector before processing. | Highly sensitive to domain shift in any modality. Struggles with temporal misalignment and creates high-dimensional, redundant representations that can amplify noise and outliers from the source domain. |
| **Late fusion** | Combines decisions from modality-specific classifiers. | More robust to completely missing modalities but ignores crucial low-level inter-modal interactions. |
| **Attention-based fusion** | Uses attention to weight features based on importance. | Can learn to attend to noisy or shifted features, amplifying errors. Performance depends on source-domain attention patterns. |
| **Distribution alignment** | Aligns heterogeneous modality distributions. | Directly aims to improve robustness but may be overfit to the alignment strategy seen in source domains. |
| **Dynamic gating** | Dynamically weights or filters modalities. | It can theoretically down-weight unreliable modalities, but the gating logic itself must adapt to new domains. |

representations that minimize the distribution shift between domains and reduce the "adaptivity gap" between a universal classifier and a domain-specific optimal one [13]. By enhancing performance on out-of-distribution (OOD) samples [8], DG strives to create highly adaptable models that bridge the gap between controlled laboratory research and robust practical application [22].

### C. TEST-TIME ADAPTATION

TTA enables pretrained models to adapt to unseen target domains during inference using only unlabeled target data, operating under a strict "source-free" constraint where original training data is unavailable. This process is inherently continual and online, as models must process data arriving in small, sequential, and potentially non-i.i.d. batches, necessitating robustness to these dynamic conditions. The primary goals of TTA are multifaceted: to achieve immediate performance enhancement on the current data stream by mitigating domain shift, while also ensuring consistent generalization to avoid over-adapting to a single batch. A critical challenge in continual settings is to prevent catastrophic forgetting, balancing adaptation with the retention of knowledge for future domains using techniques like Elastic Weight Consolidation [23]. TTA prioritizes practicality, providing a mechanism for robust adaptation without the need for labeled data or access to the source dataset [7].

## III. DG and TTA METHODS FOR MSA

To provide a structured overview, Fig. 2 presents a taxonomy of strategies for enhancing robustness in MSA. This taxonomy highlights the complementary nature of these strategies and clarifies their respective roles in addressing domain shifts and modality heterogeneity.

### A. DG APPROACHES

One family of approaches is **data manipulation**, which increases diversity during training through augmentations that simulate unseen domains. Techniques include Mixup and MixStyle [22] for feature-level blending, CutMix and RandAugment for synthetic variations, and more recently, diffusion-based generation for realistic domain shifts [8,24].

A second group is **representation learning**, which minimizes divergence between source and target feature spaces. Adversarial training aligns features via domain discriminators, while statistical

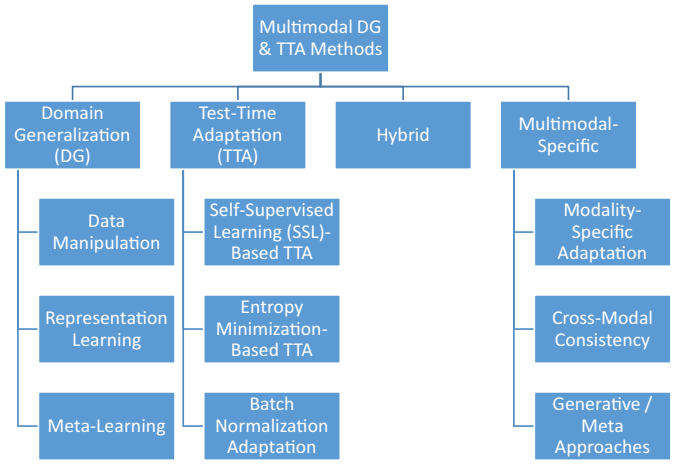



**Fig. 2.** Taxonomy of robustness strategies in MSA.

measures such as Maximum Mean Discrepancy (MMD), Correlation Alignment (CORAL) [13], and Central Moment Difference (CMD) explicitly reduce distribution gaps [25]. Disentanglement-based methods (e.g., Disentangled Sentiment Representation Adversarial Network (DiSRAN) [26]) separate style from sentiment-related features, and consistency regularization (e.g., Contrastive Adaptation and Stable Pseudo-label generation (CASP) [9], Test Time Adaptation with Transformation Invariance (TIPI)) enforces stable predictions under perturbations [10]. Architectures such as the Domain-Conditioned Transformer (DCT) embed domain cues into self-attention layers for gradual adaptation [27].

A third family is **meta-learning**, where models simulate train–test splits across source domains to develop inductive biases for rapid adaptation to unseen domains [22].

DG is closely tied to TTA, which enables real-time adaptation without source data. Self-training objectives dominate this space, including entropy minimization (e.g., Test-Time Entropy Minimization (TENT) [28]), pseudo-labeling (e.g., Source HypOthesis Transfer (SHOT) [29]), and robust variants like Efficient Anti-Forgetting Test-Time Adaptation (EATA) [30]. Other strategies include updating normalization layers, injecting domain adapters or prompts, and diffusion-based alignment (e.g., Generalized Diffusion Adaptation (GDA)) [24]. More advanced settings, such as CTTA, address streams of evolving domains while mitigating catastrophic forgetting through frameworks such as Continual test-time domain adaptation

(CoTTA) [31] and Tri-net Self-Training with Balanced Normalization (TRIBE) [32].

In summary, DG spans data augmentation, invariant representation learning, and meta-learning, while TTA complements these methods with online adaptation.

## B. TTA APPROACHES

TTA methods are often divided into fully TTA (adapting off-the-shelf models), Test-Time Training (TTT) with auxiliary self-supervised tasks, and CTTA, which tackles evolving domains while avoiding catastrophic forgetting [7].

Methods differ in regression and classification due to the nature of prediction targets and evaluation criteria. Classification focuses on assigning discrete labels, benefiting from adaptation techniques that sharpen decision boundaries and reduce prediction entropy. In contrast, regression targets continuous sentiment scores, where minor errors can significantly impact metrics like MAE. Consequently, techniques effective for classification may be less useful in regression, especially in small or noisy datasets. Regression-oriented methods should prioritize consistency-based, calibration-aware adaptations, as opposed to approaches like hard-label self-training, highlighting that entropy minimization is more effective for classification, while multimodal consistency is preferable for sentiment regression [9].

***1). SELF-SUPERVISED LEARNING (SSL)-BASED TTA.*** SSL-based TTA leverages unlabeled test data through consistency and reconstruction objectives. Contrastive methods (e.g., MixStyle-based Contrastive Test-time Adaptation (MCTTA) [22], CASP) enforce stable predictions across augmentations [9], while masked-modeling and reconstruction tasks (TTT, Meta Gradient Generator (MGG)) refine representations during inference. Pseudo-labeling frameworks such as SHOT and CoTTA extend this by filtering or averaging labels for stable continual adaptation [5,12].

***2). ENTROPY MINIMIZATION-BASED TTA.*** Pioneered by TENT, this family reduces prediction entropy to encourage confident outputs by updating batch norm parameters [28]. While effective in static settings, it struggles with small or non-i.i.d. batches. Successors such as EATA, Sharpness-Aware Robustness (SAR) [33], and Domain-specific block selection and paired-view Pseudo-Labelling for Online Test-Time adaptation (DPLOT) [11] improve robustness with redundancy filtering, sharpness-aware losses, or domain-specific block updates [7]. For regression tasks, alternatives like CASP provide entropy-free consistency.

***3). BATCH NORMALIZATION ADAPTATION.*** Batch normalization (BN)-based approaches recalibrate feature statistics on test data. Naïve Test-Time Batch Normalization (TTBN) [34,35] is unreliable in non-i.i.d. streams, but robust variants such as Robust Test-Time Adaptation (RoTTA) [36], TRIBE, and Distribution Alignment (DA-TTA) [5] use moving averages, balanced normalization, or alignment losses to stabilize adaptation [32]. Dual-BN strategies (e.g., TIPI) further enhance resilience to small batch sizes.

In summary, TTA methods range from SSL and entropy minimization to BN-based adaptation, with recent trends focusing on continual settings and multimodal regression.

## C. HYBRID AND MULTIMODAL-SPECIFIC STRATEGIES

In multimodal contexts, robustness methods often combine principles from DG with TTA mechanisms or introduce strategies tailored to the heterogeneous and noisy nature of multimodal data. Hybrid approaches, such as Training Locally Interpretable Domain Generalization Models Enables Test-time Correction (TIDE) [8], enforce domain-invariant training while enabling test-time corrections through concept-level saliency maps. Similarly, adaptive classifiers adjust decision boundaries during inference, complementing invariant representations [13]. These strategies aim to balance robustness with interpretability.

Multimodal-specific TTA directly addresses challenges like missing or corrupted modalities, modality imbalance, and the regression-oriented nature of sentiment tasks. Representative methods include CASP, which uses modality dropout and stable pseudo-labeling; Multimodal Cyclic Translation Network (MCTN) [37], which learns robust joint embeddings through cross-modal translation, and Modality-Invariant and -Specific Representations (MISA) [38], which separates invariant features from modality-specific features [39]. Generative methods such as GDA and meta-learning approaches like Meta Gradient Generator-guided TTA (MGTTA) [12] extend robustness by aligning domains or stabilizing noisy gradients, while continual frameworks (e.g., CoTTA, TRIBE) address evolving domain shifts.

A summary of representative strategies is provided in Table III, highlighting the spectrum of DG, TTA, hybrid, and multimodal-specific methods, their core mechanisms, best setting, main strength, and limitation.

# IV. REAL-WORLD CHALLENGES AND DEPLOYMENT ISSUES

## A. UNSEEN DATA DISTRIBUTION

Real-world data streams present a specific form of distribution shift by violating i.i.d. assumptions. Unlike static domain shifts seen in laboratory benchmarks, these non-i.i.d. streams introduce local biases and temporal correlations that lead to conflicting optimization objectives during test-time [5]. These shifts manifest across all modalities: changes in speaker demographics, background audio noise, visual variabilities (lighting, resolution, viewpoint), and linguistic shifts like slang, colloquialisms, and errors from ASR systems [9,40,41]. Furthermore, the inherent heterogeneity between modalities, where text is high-level and semantic, while audio and visual features are lower-level, creates significant distribution disparities [42]. This complex reality of non-i.i.d., continually shifting data is the primary challenge that TTA and robust multimodal learning strategies aim to address.

## B. MULTIMODAL-SPECIFIC CHALLENGES

MSA encounters unique challenges that differ from unimodal settings, primarily due to varying reliability, timing, and semantic roles of input streams. Key issues include **modality imbalance**, where text often dominates sentiment signals while audio and visual cues may be weaker or noisier [43,44]. **Cross-modal inconsistency** arises when conflicting cues, like positive words coupled with a negative tone, confuse fusion mechanisms, leading to unstable predictions [17]. Additionally, **alignment errors** between audio, text, and visual streams complicate the correct mapping of evidence to words, significantly impacting sentiment prediction accuracy [19]. These issues are exacerbated by frequent **missing or corrupted modalities**, necessitating robust models that can recover or adapt under such conditions [25]. Overall, enhancing multimodal robustness requires distinct strategies to address

**Table III.** Comparison of DG, TTA, and hybrid multimodal strategies in MSA

| Category | Methods | Core mechanism | Best setting | Main strength | Main limitation |
|---|---|---|---|---|---|
| **DG** | MixStyle, CORAL | Learning domain-invariant features via data augmentation, divergence minimization, or disentangled representations. | Classification and regression. | Lightweight and effective for learning domain-invariant features. | No online adaptation after deployment. |
| **TTA—SSL-based** | CASP, MCTTA | Contrastive learning, modality dropout, and transformation invariance guide adaptation using only target data. | Classification and regression. | Learn modality-invariant features that are highly robust to environmental noise. | Improper tuning of stability threshold and internal parameters can either include too much noise or too few samples for effective self-training. |
| **TTA—entropy-based** | TENT, EATA, DPLOT | Minimizing prediction entropy, adapting BN or domain-specific blocks, and filtering noisy samples. | Classification. | Simple and effective for fast test-time adaptation. | Entropy-based updates can be unstable on small or highly non-i.i.d. batches. |
| **TTA—BN adaptation** | RoTTA, TRIBE, DA-TTA | Updating or balancing BN statistics/affine layers with test streams; align distributions. | Classification and regression. | Highly efficient for real-time edge deployment as it primarily involves updating running statistics. | Naive BN adaptation is unreliable in dynamic online data streams or under global class imbalance, where statistics become biased toward the most frequent incoming class. |
| **Hybrid** | TIDE, Adaptive Classifiers. | DG (domain-invariant training) combined with test-time correction via saliency maps or adaptive decision boundaries. | Classification. | Balances robustness with interpretability and flexibility. | Often more complex to design and tune than single-paradigm methods. |
| **Multimodal-specific** | MCTN, MISA | Addressing noisy/missing modalities via dropout, cross-modal translation, or invariant+specific feature disentanglement. | Classification and regression. | Directly addresses multimodal heterogeneity. | Performance depends on how well cross-modal structure is learned. |
| **Generative/Meta** | GDA, MGTTA, CoTTA | Diffusion-based domain alignment, meta-learned optimizers, or continual TTA with weight averaging. | Classification. | Promising for complex shifts and stronger adaptation stability. | More computationally expensive and less mature than simpler methods. |

structural weaknesses inherent in multimodal inputs, separate from general domain adaptation and transfer techniques.

## C. MISSING MODALITIES

The absence or severe corruption of one or more data streams is a frequent occurrence in real-world deployments, such as a voice-only call lacking video or noisy audio, which can become unreliable. This presents a major challenge, as standard models assume complete and clean input, leading to significant performance degradation when modalities are missing. To combat this, strategies focus on building inherent robustness and explicit recovery. Architectures like the MCTN are designed to be inherently robust, learning to translate between modalities so that inference requires only a single source modality [37]. Similarly, MISA learns two distinct representations—a modality-invariant representation and a modality-specific one. Explicit techniques include reconstruction and imputation (e.g., Multimodal Sentiment Analysis Missing Modality Reconstruction Network Based on Shared-Specific Features (SSF-MMRN) [25] reconstructs missing modalities and checks for semantic consistency) and consistency enforcement (e.g., CASP's modality random dropout trains the model to be consistent across different modality subsets) [9]. Finally, adaptive fusion mechanisms, such as the dynamic gating in Dynamic Weighted Gating for Enhanced Cross-Modal Interaction (DWGCI) [19] or the adaptive weighting in Multimodal Transformer with Adaptive Modality Weighting (MTAMW) [39], allow the model to dynamically down-weight the influence of a missing or unreliable modality during fusion.

## D. COMPUTATIONAL AND LATENCY CONSTRAINTS

Deploying models on edge devices and in real-time applications imposes strict constraints on computational complexity, memory footprint, and latency. High-performance servers are often unavailable, and models must operate within the limited processing power, memory (VRAM), and battery life of edge devices. Applications

such as autonomous driving and real-time social media monitoring require high-speed processing and immediate adaptation with low latency, often processing data in small sequential batches [10]. This necessitates computationally lightweight methods. Strategies to achieve this include parameter-efficient transfer learning (updating only a small subset of parameters, like normalization layers), feature optimization to remove redundancies [27,45], and the use of efficient architectures such as ALBERT [46] or ELECTRA [47]. Designing efficient fusion mechanisms that avoid quadratic complexity is also critical. Furthermore, methods must demonstrate fast convergence and robustness to the instability caused by small batch sizes to be practical in these constrained environments.

### E. PRIVACY AND SECURITY

The "source-free" nature of many TTA methods provides a significant advantage for privacy-sensitive applications. A key reason source data is unavailable during deployment is due to privacy regulations and business confidentiality concerns, as training data often contains sensitive information [7]. TTA circumvents this by adapting a pretrained model using only the incoming stream of unlabeled target data, eliminating the need to store or access the source dataset during inference. This addresses privacy concerns and reduces storage overhead on deployment devices. This makes TTA particularly suitable for domains such as healthcare, where a model trained on a public dataset can adapt to a new hospital scanner or a specific patient's data without ever requiring access to the original, potentially sensitive, training records [48].

## V. DATASETS AND EVALUATION

### A. COMMON MSA DATASETS

Benchmark MSA datasets are essential for evaluating model performance under DG and TTA scenarios. These datasets provide multimodal content (text, audio, visual) annotated with sentiment labels, often simulating real-world challenges.

Key datasets include the **CMU-MOSI** and **CMU-MOSEI** datasets, which are the primary benchmarks for English MSA. They contain video clips of movie reviews with sentiment annotations on a scale from -3 (strongly negative) to +3 (strongly positive) [49]. MOSEI is larger and more diverse, making it crucial for assessing generalization [50]. The **CH-SIMS** dataset provides a Chinese counterpart with fine-grained unimodal and multimodal labels, enabling research into cross-lingual domain shifts [51]. To specifically test robustness to real-world noise, the **MOSI-ASR** variants (e.g., MOSI-SpeechBrain) are created by processing MOSI audio with ASR systems, replacing clean transcripts with error-prone ASR output. This directly introduces a distribution shift for the textual modality [52].

Other datasets are used for cross-domain evaluation. **MVSA**, comprising image–text pairs from Twitter [53], and **Yelp** review data are paired in cross-domain sentiment classification tasks (e.g., MVSA ↔ Yelp) [26,54]. For complex affective computing, the **Video-SA** dataset includes contexts like sarcasm and metaphor [55]. While not exclusively for sentiment, datasets like **IEMOCAP** (emotion recognition) [56], **UR-FUNNY** (humor detection) [57], and **VSO** (a pretraining dataset) are also utilized to evaluate the generalization of multimodal affective models [58]. A summary of these datasets is provided in Table IV.

### B. EVALUATION METRICS

Evaluating MSA models, particularly for DG and TTA, requires a suite of metrics to measure both standard performance and robustness under domain shift. These metrics are categorized based on the primary task (classification or regression) and the specific challenge being evaluated.

***1). CORE PERFORMANCE METRICS.*** Performance is measured differently for classification and regression tasks, as summarized in Table V.

For classification tasks, accuracy is commonly measured for different numbers of classes: Acc-2 (binary), Acc-3 (ternary), Acc-5, and Acc-7 [1]. The F1-score, the harmonic mean of precision and recall, is a crucial supplementary metric for imbalanced datasets. For regression tasks predicting sentiment intensity, the standard metrics are MAE and Pearson's Correlation (Corr), which measure prediction error and alignment with true labels, respectively [17].

***2). ROBUSTNESS METRICS.*** Beyond core performance, assessing model behavior under challenging conditions is paramount for DG and TTA. Robustness metrics assess model performance under domain shift and non-ideal conditions.

**Table IV.** Commonly used datasets in MSA

| Dataset | Modalities | Typical use in DG/TTA/robustness |
|---|---|---|
| **CMU-MOSI** | Text, Audio, Visual | Benchmark for sentiment regression/classification; often used in TTA studies. |
| **CMU-MOSEI** | Text, Audio, Visual | Large-scale benchmark with diverse speakers; widely used for generalization and DG. |
| **CH-SIMS** | Text, Audio, Visual (Chinese) | Cross-lingual evaluation; fine-grained sentiment labels. |
| **MVSA-Single/multiple** | Image, Text | Used in multimodal cross-domain sentiment classification (e.g., MVSA ↔ Yelp). |
| **MOSI-ASR (SpeechBrain, IBM, iFlytek)** | Text (ASR), Audio, Visual | Tests robustness against ASR transcription errors. |
| **IEMOCAP** | Text, Audio, Visual | Emotion recognition dataset adapted for sentiment-related DG tasks. |
| **Yelp** | Image, Text | Paired with MVSA for cross-domain generalization studies. |
| **Video-SA** | Text, Audio, Visual | Evaluates models under complex contexts (sarcasm, metaphor). |
| **UR-FUNNY/MUStARD** [59] | Text, Audio, Visual | Humor detection datasets, often extended to affective MSA. |
| **VSO (Visual Sentiment Ontology)** | Image, Text | Pretraining resource for sentiment-aware multimodal models. |

**Table V.** Summary of core evaluation metrics for MSA

| Metric | Description and use case | Task |
|---|---|---|
| **Acc-2/F1** | Accuracy and F1-score for binary classification (e.g., Negative vs. Non-Negative). F1 is preferred for imbalanced datasets. | Classification |
| **Acc-3/F1** | Accuracy and F1-score for 3-class classification (Positive, Neutral, Negative). | Classification |
| **Acc-5/Acc-7** | Accuracy for finer-grained sentiment classification (e.g., strongly negative to strongly positive). | Classification |
| **Mean absolute error (MAE)** | Measures the average absolute prediction error. Primary metric for sentiment intensity regression. | Regression |
| **Pearson correlation (Corr)** | Measures the linear correlation between predictions and true labels. Assesses how well prediction trends match ground truth. | Regression |

This involves evaluating on corrupted datasets like MOSI-ASR or standard corruption benchmarks (e.g., CIFAR-C), reporting the standard performance metrics from Table V under noise to quantify the performance drop [7]. For TTA, stability with small batch sizes is a key benchmark, as many methods fail in this setting. In CTTA scenarios, the average performance over a sequence of changing domains is measured, focusing on the ability to prevent catastrophic forgetting and error accumulation [10,23]. Finally, ablation studies are routinely conducted to quantify the contribution of specific components (e.g., a module designed to correct ASR errors) to the overall robustness and performance of the model [18,40].

# VI. OPEN CHALLENGES AND FUTURE DIRECTIONS

## A. MANAGING MIXED AND CONTINUOUS SHIFTS

Real-world environments are dynamic and unpredictable, presenting models with a sequence of evolving and overlapping challenges—mixed and continual shifts—that severely degrade performance. Furthermore, shifts are rarely isolated; they involve combinations of corruptions (e.g., ASR errors + visual noise + linguistic slang) simultaneously [12]. Traditional TTA fails here for three reasons: batch size instability, catastrophic forgetting of past knowledge, and error accumulation from noisy labels [7,10,11]. The paramount challenge is to develop methods that balance the enhancement of performance on the current domain (without labels) with the need to avoid excessive bias, thereby ensuring long-term generalization across a stream of unseen domains.

## B. THE ROLE OF LARGE MULTIMODAL MODELS (LMMs)

Large multimodal models (LMMs)—including transformer-based language models such as BERT and multimodal encoders like CLIP [60]—have become key components in modern MSA pipelines. They are adapted by using their powerful encoders for feature extraction (e.g., BERT for text, CLIP-ViT for visuals) and fine-tuned with sophisticated strategies, such as the two-phase approach in Two-Phase Multi-task Sentiment Analysis (TPMSA), which first optimizes BERT on text before integrating other modalities [61]. However, DG and TTA for these models differ significantly from traditional methods. The focus shifts to parameter-efficient and source-free adaptation, where only a minimal subset of parameters (e.g., normalization layers, conditioning vectors) is updated online to avoid the prohibitive cost of full retraining [27]. This has spurred novel paradigms like using diffusion models (e.g., GDA) to generate domain-aligned samples without changing weights and meta-learning optimizers (e.g., MGTTA) that learn to refine noisy gradients during TTA. Adapting these frameworks for regression is a major challenge, as most existing solutions focus only on classification [9].

## C. EXPLAINABILITY AND INTERPRETABILITY

The "black-box" nature of deep learning models is a significant barrier to their deployment in high-stakes applications. This challenge is amplified in multimodal settings due to the heterogeneity of data and the complexity of fusion strategies. Understanding why a model fails is crucial for debugging, ensuring fairness, complying with regulations, and building user trust. Research is addressing this through several avenues: Visualization techniques such as attention maps (e.g., Attention Rollout) and saliency maps (e.g., Grad-CAM) to highlight which parts of the input the model focuses on [24,27]. Concept-level interpretability frameworks like TIDE train models to rely on local, human-understandable concepts (e.g., "*beak*" and "*feathers*") and use these for both explanation and even test-time correction [8]. Quantitative analyses, such as ablation studies and measuring modality contributions via gating mechanisms, provide insight into which components are most critical for robust performance. The goal is to move beyond mere performance metrics and provide granular explanations for model behavior in dynamic environments [62].

## D. BENCHMARKING AND STANDARDIZATION

The rapid development of new DG and TTA methods necessitates standardized protocols for fair and objective comparison, as the current lack of uniformity in experimental setups hinders the assessment of true progress. First, the field requires standardized datasets that accurately reflect real-world complexities, such as class imbalance [32], MOSI-ASR variants for simulating ASR errors, and large-scale, multi-domain benchmarks like Geo-YFCC [13]. Second, evaluation metrics must extend beyond basic accuracy to include robust measures assessed across both clean and corrupted inputs. This involves utilizing F1-score and accuracy for classification, MAE and Pearson's correlation for regression, and specific benchmarks for non-i.i.d. streams and continual shifts. Finally, future studies should implement rigorous, uniform experimental protocols [7]. These must mandate fixed source-target splits, shared adaptation budgets (ensuring consistent batch sizes and update steps) [12], detailed ablation studies, and comparisons against a common set of strong baselines [9] reported across multiple random seeds for reliability [17]. Establishing these comprehensive standards is indispensable for improving the transparency, robustness, and long-term progress of MSA models.

# VII. CONCLUSION

In conclusion, this review highlighted the critical significance of TTA and DG in enhancing the practical deployment of MSA

systems. As MSA models transition from controlled laboratory settings to dynamic real-world environments, they face substantial challenges, primarily due to distribution shifts that can severely impact performance. The comprehensive survey presented here categorized existing TTA and DG methodologies, offering insights into their strengths and limitations while identifying key datasets and evaluation metrics vital for assessing model robustness.

The findings underscored the necessity for advanced strategies that could effectively address the complexities of mixed and continual domain shifts. This ensures that models maintain high performance across diverse and evolving data distributions. Future research avenues are promising, particularly the exploration of hybrid approaches that integrate the strengths of both TTA and DG, as well as the adaptation of LMMs to leverage their extensive pretrained knowledge. Moreover, the emphasis on explainability and interpretability remains paramount, as understanding the decision-making processes of these complex models is crucial for fostering trust and reliability in applications where sentiment analysis plays a pivotal role.

By prioritizing these innovative research directions, the field can move closer to achieving effective sentiment analysis systems that accurately capture the intricacies of human emotional expression in real-world contexts. This will be crucial for developing MSA systems that are not only accurate but also truly trustworthy and reliable tools in our daily interactions with technology.

## ACKNOWLEDGMENTS

This research was supported by the Faculty of Computing and Information Technology, Tunku Abdul Rahman University of Management and Technology (TAR UMT), Malaysia. Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the supporting institution.

## CONFLICT OF INTEREST STATEMENT

The authors declare that they have no conflicts of interest to report regarding the present study.

## REFERENCES

[1] Y. Li, A. Liu, and Y. Lu, "Multi-level language interaction transformer for multimodal sentiment analysis," *J. Intell. Inf. Syst.*, vol. 63, no. 3, pp. 945–964, Feb. 2025, doi: 10.1007/s10844-025-00923-x.

[2] J. Peng *et al.*, "A fine-grained modal label-based multi-stage network for multimodal sentiment analysis," *Expert Syst. Appl.*, vol. 221, p. 119721, Jul. 2023, doi: 10.1016/j.eswa.2023.119721.

[3] H. Cheng *et al.*, "Multimodal sentiment analysis based on attentional temporal convolutional network and multi-layer feature fusion," *IEEE Trans. Affect. Comput.*, vol. 14, no. 4, pp. 3149–3163, Apr. 2023, doi: 10.1109/TAFFC.2023.3265653.

[4] J. Li and H. Zhao, "Multimodal Sentiment Analysis Method Based on Multi-task Learning," in *2023 6th International Conference on Signal Processing and Machine Learning (SPML)*, New York, NY, USA: ACM, Oct. 2023, pp. 308–314. doi: 10.1145/3614008.3614055.

[5] Z. Wang *et al.*, "Distribution alignment for fully test-time adaptation with dynamic online data streams," in *Lect. Notes Comput. Sci.*, pp. 332–349, 2024, doi: 10.1007/978-3-031-72691-0_19.

[6] D. Patel and K. Amin, "A cross-domain semantic similarity measure and multi-source domain adaptation in sentiment analysis," in *2022 International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, IEEE, Nov. 2022, pp. 760–764. doi: 10.1109/ICAISS55157.2022.10011051.

[7] X. Yang *et al.*, "A versatile framework for continual test-time domain adaptation: Balancing discriminability and generalizability," *2024 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2024, pp. 23731–23740. doi: 10.1109/CVPR52733.2024.02240.

[8] A. Agarwal, S. Karanam, and V. Gandhi, "TIDE: Training locally interpretable domain generalization models enables test-time correction," *2025 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2025, pp. 30210–30220. doi: 10.1109/CVPR52734.2025.02812.

[9] Z. Guo *et al.*, "Bridging the gap for test-time multimodal sentiment analysis," Proc. AAAI Conf. Artif. Intell., vol. 39, no. 16, pp. 16987–16995, Apr. 2025, doi: 10.1609/aaai.v39i16.33867.

[10] A. T. Nguyen *et al.*, "TIPI: Test time adaptation with transformation invariance," *2023 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2023, pp. 24162–24171. doi: 10.1109/CVPR52729.2023.02314.

[11] Y. Yu *et al.*, "Domain-specific block selection and paired-view pseudo-labeling for online test-time adaptation," *2024 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2024, pp. 22723–22732. doi: 10.1109/CVPR52733.2024.02144.

[12] Q. Deng *et al.*, "Learning to generate gradients for test-time adaptation via test-time training layers," Proc. AAAI Conf. Artif. Intell., vol. 39, no. 15, pp. 16235–16243, Apr. 2025, doi: 10.1609/aaai.v39i15.33783.

[13] A. Dubey *et al.*, "Adaptive methods for real-world domain generalization," *2021 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2021, pp. 14335–14344. doi: 10.1109/CVPR46437.2021.01411.

[14] J. Devlin *et al.*, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North, Stroudsburg, PA, USA: Association for Computational Linguistics*, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[15] S. Schneider *et al.*, "wav2vec: Unsupervised pre-training for speech recognition," *Interspeech 2019, ISCA: ISCA, Sep.* 2019, pp. 3465–3469. doi: 10.21437/Interspeech.2019-1873.

[16] K. He *et al.*, "Deep residual learning for image recognition," *2016 IEEE Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[17] W. Geng, X. Li, and Y. Bian, "A dual-branch enhanced multi-task learning network for multimodal sentiment analysis," in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, New York, NY, USA: ACM, Jun. 2023, pp. 481–489. doi: 10.1145/3591106.3592260.

[18] J. Huang *et al.*, "Multimodal sentiment analysis in realistic environments based on cross-modal hierarchical fusion network," *Electronics (Basel).*, vol. 12, no. 16, p. 3504, Aug. 2023, doi: 10.3390/electronics12163504.

[19] N. Wang and Q. Wang, "Dynamic weighted gating for enhanced cross-modal interaction in multimodal sentiment analysis," *ACM Trans. Multimed. Comput. Commun. Appl.*, vol. 21, no. 1, pp. 1–19, Jan. 2025, doi: 10.1145/3702996.

[20] A. Yadav and D. K. Vishwakarma, "A deep multi-level attentive network for multimodal sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 1, pp. 1–19, Jan. 2023, doi: 10.1145/3517139.

[21] Y. Zhang *et al.*, "Multimodal sentiment analysis network based on distributional transformation and gated cross-modal fusion," *in 2024 International Conference on Networking and Network Applications (NaNA)*, IEEE, Aug. 2024, pp. 496–503. doi: 10.1109/NaNA63151.2024.00088.

[22] K. Yamashita and K. Hotta, "MixStyle-based contrastive test-time adaptation: Pathway to domain generalization," *2024 IEEE/CVF Conf. Comput. Vision Pattern Recogn. Workshops (CVPRW)*, IEEE, Jun. 2024, pp. 1029–1037. doi: 10.1109/CVPRW63382.2024.00109.

[23] A. Madasu and A. R. Vijjini, "Sequential domain adaptation through elastic weight consolidation for sentiment analysis," *in 2020 25th International Conference on Pattern Recognition (ICPR)*, IEEE, Jan. 2021, pp. 4879–4886. doi: 10.1109/ICPR48806.2021.9412617.

[24] Y.-Y. Tsai *et al.*, "GDA: Generalized diffusion for robust test-time adaptation," *2024 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2024, pp. 23242–23251. doi: 10.1109/CVPR52733.2024.02193.

[25] Y. Guo *et al.*, "Multimodal sentiment analysis missing modality reconstruction network based on shared-specific features," in *2023 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*, IEEE, Nov. 2023, pp. 1–6. doi: 10.1109/ICSMD60522.2023.10490667.

[26] Y. Zhang *et al.*, "Learning disentangled representation for multimodal cross-domain sentiment analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7956–7966, Feb. 2022, doi: 10.1109/TNNLS.2022.3147546.

[27] Y. Tang *et al.*, "Domain-conditioned transformer for fully test-time adaptation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2024, pp. 6260–6269. doi: 10.1145/3664647.3680678.

[28] D. Wang *et al.*, "Tent: Fully test-time adaptation by entropy minimization," in *International Conference on Learning Representations*, Mar. 2021.

[29] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation," in *ICML'20: Proceedings of the 37th International Conference on Machine Learning*, Jul. 2020.

[30] S. Niu *et al.*, "Efficient test-time model adaptation without forgetting," in *International Conference on Machine Learning*, May 2022.

[31] Q. Wang *et al.*, "Continual test-time domain adaptation," *2022 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2022, pp. 7191–7201. doi: 10.1109/CVPR52688.2022.00706.

[32] Y. Su, X. Xu, and K. Jia, "Towards real-world test-time adaptation: Tri-net self-training with balanced normalization," Proc. AAAI Conf. Artif. Intell., vol. 38, no. 13, pp. 15126–15135, Mar. 2024, doi: 10.1609/aaai.v38i13.29435.

[33] S. Niu *et al.*, "Towards stable test-time adaptation in dynamic wild world," in *International Conference on Learning Representations (ICLR) 2023*, Feb. 2023.

[34] Z. Nado *et al.*, "Evaluating prediction-time batch normalization for robustness under covariate shift," ArXiv, Jan. 2021.

[35] S. Schneider *et al.*, "Improving robustness against common corruptions by covariate shift adaptation," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Oct. 2020.

[36] L. Yuan, B. Xie, and S. Li, "Robust test-time adaptation in dynamic scenarios," *2023 IEEE/CVF Conf. Comput. Vision Pattern Recogn. (CVPR)*, IEEE, Jun. 2023, pp. 15922–15932. doi: 10.1109/CVPR52729.2023.01528.

[37] H. Pham *et al.*, "Found in translation: Learning robust joint representations by cyclic translations between modalities," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 01, pp. 6892–6899, Jul. 2019, doi: 10.1609/aaai.v33i01.33016892.

[38] D. Hazarika, R. Zimmermann, and S. Poria, "MISA: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, New York, NY, USA: ACM, Oct. 2020, pp. 1122–1131. doi: 10.1145/3394171.3413678.

[39] Y. Wang *et al.*, "Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis," *Neurocomputing*, vol. 572, p. 127181, Mar. 2024, doi: 10.1016/j.neucom.2023.127181.

[40] J. Huang *et al.*, "Heterogeneous hierarchical fusion network for multimodal sentiment analysis in real-world environments," *Electronics (Basel).*, vol. 13, no. 20, p. 4137, Oct. 2024, doi: 10.3390/electronics13204137.

[41] E. Veltmeijer and C. Gerritsen, "Automatic domain-adaptive sentiment analysis with SentiMap," *Int. J. Semant. Comput.*, vol. 18, no. 01, pp. 97–120, Mar. 2024, doi: 10.1142/S1793351X24410058.

[42] Y. Zheng *et al.*, "DJMF: A discriminative joint multi-task framework for multimodal sentiment analysis based on intra- and inter-task dynamics," *Expert Syst. Appl.*, vol. 242, p. 122728, May 2024, doi: 10.1016/j.eswa.2023.122728.

[43] D. Gkoumas *et al.*, "What makes the difference? An empirical comparison of fusion strategies for multimodal language analysis," *Inf. Fusion*, vol. 66, pp. 184–197, Feb. 2021, doi: 10.1016/j.inffus.2020.09.005.

[44] W. Han, H. Chen, and S. Poria, "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*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 9180–9192. doi: 10.18653/v1/2021.emnlp-main.723.

[45] C. Ding *et al.*, "Multimodal sentiment analysis via efficient multimodal transformer and modality-aware adaptive training strategy," in *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, New York, NY, USA: ACM, Oct. 2023, pp. 11–17. doi: 10.1145/3606039.3613113.

[46] Z. Lan *et al.*, "ALBERT: A lite BERT for self-supervised learning of language representations," in *International Conference on Learning Representations*, Feb. 2020.

[47] K. Clark *et al.*, "ELECTRA: Pre-training text encoders as discriminators rather than generators," in *8th International Conference on Learning Representations*, Mar. 2020.

[48] T. Varsavsky *et al.*, "Test-time unsupervised domain adaptation," in Lecture Notes in Computer Science, vol. 12261, pp. 428–436, 2020, doi: 10.1007/978-3-030-59710-8_42.

[49] A. Zadeh *et al.*, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intell. Syst.*, vol. 31, no. 6, pp. 82–88, Nov. 2016, doi: 10.1109/MIS.2016.94.

[50] A. Bagher Zadeh *et al.*, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 2236–2246. doi: 10.18653/v1/P18-1208.

[51] W. Yu *et al.*, "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 3718–3727. doi: 10.18653/v1/2020.acl-main.343.

[52] Y. Wu *et al.*, "Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR Errors," in Findings of the Association for Computational Linguistics: ACL *2022*, Stroudsburg, PA, USA: Association for Computational Linguistics, Mar. 2022, pp. 1397–1406. doi: 10.18653/v1/2022.findings-acl.109.

[53] T. Niu *et al.*, "Sentiment analysis on multi-view social data," in Lecture Notes in Computer Science, vol. 9517, pp. 15–27, 2016, doi: 10.1007/978-3-319-27674-8_2.

[54] Q.-T. Truong and H. W. Lauw, "VistaNet: Visual aspect attention network for multimodal sentiment analysis," Proc. AAAI Conf. Artif.

Intell., vol. 33, no. 01, pp. 305–312, Jul. 2019, doi: 10.1609/aaai.v33i01.3301305.

[55] J. Wang *et al.*, "AMSA: Adaptive multimodal learning for sentiment analysis," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 19, no. 3s, pp. 1–21, Feb. 2023, doi: 10.1145/3572915.

[56] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, Feb. 2008, doi: 10.1007/s10579-008-9076-6.

[57] M. K. Hasan *et al.*, "UR-FUNNY: A multimodal language dataset for understanding humor," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2019, pp. 2046–2056. doi: 10.18653/v1/D19-1211.

[58] J. Ye *et al.*, "Sentiment-aware multimodal pre-training for multimodal sentiment analysis," *Knowl. Based. Syst.*, vol. 258, p. 110021, Dec. 2022, doi: 10.1016/j.knosys.2022.110021.

[59] S. Castro *et al.*, "Towards multimodal sarcasm detection (An obviously perfect paper)," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics, Jul. 2019, pp. 4619–4629. doi: 10.18653/v1/P19-1455.

[60] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, Feb. 2021.

[61] B. Yang *et al.*, "Multimodal sentiment analysis with two-phase multi-task learning," IEEE/ACM Trans. Audio Speech Lang. Process., vol. 30, pp. 2015–2024, 2022, doi: 10.1109/TASLP.2022.3178204.

[62] A. Ando *et al.*, "On the use of modality-specific large-scale pre-trained encoders for multimodal sentiment analysis," *2022 IEEE Spoken Lang. Technol. Workshop (SLT)*, IEEE, Jan. 2023, pp. 739–746. doi: 10.1109/SLT54892.2023.10022548.