Generative-to-Discriminative Test-Time Adaptation via Manifold-Aware Diffusion and Bayesian Distillation

Boyun Zhang[1,*] **Zequn Xie**[1,*] **Fangming Feng**[1,*] **Zihan Zhang**[1] **Yongbo He**[1]
Chuxin Wang[1] **Sihang Cai**[1] **Tao Jin**[1] **Qifei Zhang**[1,†]
[1]Zhejiang University
[*]Equal contribution
[†]Corresponding author
Contact email: boyun-zhang@zju.edu.cn

Abstract

Multimodal Sentiment Analysis (MSA) models typically suffer significant performance degradation under domain shifts. While Test-Time Adaptation (TTA) aims to mitigate this, existing discriminative approaches often succumb to "confident but wrong" predictions on out-of-distribution samples. Conversely, generative models offer robust calibration but incur prohibitive computational costs. To bridge this gap, we propose **GD-Adapt** (Generative-Discriminative Adaptation), a novel TTA framework that harmonizes the robustness of generative diffusion models with the efficiency of discriminative regression networks via **Bayesian Diffusion Distillation (BDD)**. Specifically, we introduce **Auxiliary Generative Regularization (AGR)** during pretraining to enforce manifold-aware feature learning. Extensive experiments across five cross-domain scenarios demonstrate our method's superiority. For instance, on the challenging MOSI → SIMS shift, GD-Adapt reduces Mean Absolute Error (MAE) from 0.6872 to **0.5673** and boosts binary accuracy by **5.81 percentage points** (reaching **57.33%**). Notably, in scenarios such as SIMS → MOSI, we achieve an **11.18-point** gain over the non-adapted baseline.

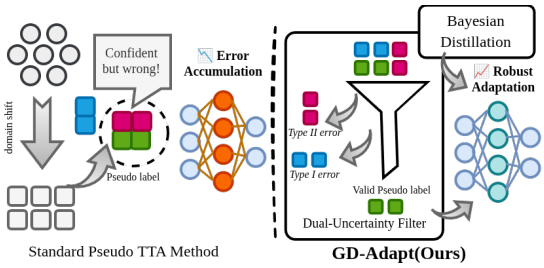

Figure 1: **Conceptual comparison between Standard TTA and our GD-Adapt. Left:** Traditional discriminative Pseudo-Labeling TTA methods often suffer from the "confident but wrong" failure mode, blindly adapting to OOD noise (Type II errors) leading to error accumulation. **Right:** GD-Adapt employs a generic **Dual-Uncertainty Filter** (visualized as the funnel) to explicitly reject unstable and OOD samples. We then perform **Bayesian Distillation**, allowing a fast student model to inherit robust uncertainty calibration from a generative teacher online.

1 Introduction

Multimodal Sentiment Analysis (MSA) has witnessed significant strides with the advent of large-scale pretraining and sophisticated fusion architectures (Wang et al., 2025a). However, state-of-the-art models typically operate under the i.i.d. assumption. In real-world deployments—ranging from social media monitoring to real-time interaction systems—models inevitably encounter distribution shifts caused by sensor noise, environmental changes, or linguistic drift. Similar robustness and generalization challenges have also been observed in other multimodal settings, such as cross-domain video understanding and retrieval (Feng et al., 2026; Xie et al., 2026). When a model trained on source data is deployed in such dynamic target environments, its performance often degrades catastrophically, a phenomenon known as the *robustness gap*.

To mitigate this, **Test-Time Adaptation (TTA)** has emerged as a practical paradigm, allowing models to adapt to streaming unlabeled target data during inference (Wang et al., 2021, 2025b; Sun et al., 2020). Standard TTA methods, such as Entropy Minimization (Wang et al., 2021) and Pseudo-Labeling (Liang et al., 2021), have shown promise but suffer from a critical limitation: they are fundamentally discriminative and lack awareness of the underlying data manifold. These methods blindly reinforce predictions where the model is confident, assuming that high confidence equates to correctness. However, recent studies (Niu et al., 2023; Yuan et al., 2023) highlight that under severe shifts, discriminative models succumb to *Type II errors*—they become "confident but wrong." This

Findings of the Association for Computational Linguistics: ACL 2026, pages 7052–7062
July 2-7, 2026

leads to error accumulation and devastating model collapse (as illustrated in Figure 1, Left). Conversely, generative models, particularly Diffusion Probabilistic Models (Ho et al., 2020; Gao et al., 2023), offer superior density estimation and calibration. Yet, their iterative denoising process incurs prohibitive computational latency, rendering them impractical for low-latency online adaptation (Prabhudesai et al., 2023).

In this paper, we bridge the gap between *generative robustness* and *discriminative efficiency*. We argue that robustness requires learning the underlying data manifold rather than just decision boundaries—a concept we term the *Paradox of Robustness*, where superficial source accuracy is often sacrificed for structural validity during pretraining. We propose **GD-Adapt** (Generative-Discriminative Adaptation), a novel framework that leverages a robust generative "Teacher" to guide a lightweight discriminative "Student" via **Bayesian Diffusion Distillation (BDD)**.

GD-Adapt is founded on three pillars. First, we introduce **Auxiliary Generative Regularization (AGR)** during pretraining. By coupling standard regression with a diffusion-based denoising task and **Stochastic Modality Perturbation (SMP)**, we force the shared encoder to capture semantically rich features rather than shortcut correlations. Second, to address the risk of adapting to out-of-distribution (OOD) noise, we introduce a **Source Manifold Estimator (SME)** based on Normalizing Flows. The SME acts as a "Manifold Gatekeeper," calculating the exact log-likelihood of incoming test samples to reject inputs that violate the source density support.

Finally, during the adaptation phase, we freeze the generative head and employ it as a Bayesian Teacher. Through ensemble sampling, the teacher provides uncertainty-calibrated pseudo-labels to update the fast regression head (the Student). This **Dual-Uncertainty Filter (DUF)**—combining epistemic uncertainty estimates from the Flow module and aleatoric stability from the Diffusion ensemble—ensures that the model adapts only on structurally valid and stable samples (Figure 1, Right).

Our contributions are summarized as follows:

- We propose the generic **GD-Adapt** architecture that harmonizes generative density modeling with discriminative inference speed, solving the latency bottleneck of existing generative TTA approaches (Gao et al., 2023).
- We introduce the **Source Manifold Estimator (SME)**, a flow-based module that explicitly detects manifold shifts, effectively mitigating the "confident but wrong" failure mode prevalent in entropy-based TTA (Wang et al., 2021; Niu et al., 2023).
- We develop a **Bayesian Diffusion Distillation** strategy that allows a regression student to inherit the uncertainty calibration of a diffusion teacher online. Experiments demonstrate that GD-Adapt recovers performance on shifted domains while maintaining low-latency Student-side inference.

2 Related Work

2.1 Multimodal Sentiment Analysis under Shifts

Sentiment analysis has evolved from early lexicon-based methods(Baccianella et al., 2010; Kiritchenko et al., 2014) to sophisticated deep learning architectures. To capture the full spectrum of human emotion, **Multimodal Sentiment Analysis (MSA)** incorporates acoustic and visual signals alongside text (Soleymani et al., 2017), leveraging complex fusion strategies to enhance performance (Morency et al., 2011). Recent advancements utilize adaptive attention (Wang et al., 2025a) and disentangled representation learning (Li et al., 2025b) to bridge semantic gaps. Despite these strides, most MSA frameworks operate under the assumption that training and testing data are independent and identically distributed (i.i.d.). In real-world deployments, non-stationary environments and sensor noise introduce distribution shifts, causing severe performance degradation (Meng et al., 2019). While domain adaptation has been widely studied, it typically assumes a fixed target domain available in batches, ignoring the continuously evolving nature of online data streams (Hoffman et al., 2014).

2.2 Test-Time Adaptation (TTA)

Test-Time Adaptation (TTA) addresses shifts by adapting a pre-trained model to unlabeled target data online (Sun et al., 2020; Liang et al., 2025). Unlike Domain Generalization (DG) (Zhou et al., 2023), which freezes the model after training, TTA continuously updates parameters during inference. Early methods like Tent (Wang et al., 2021) minimize prediction entropy to update normalization layers. However, in dynamic wild scenarios, naive

entropy minimization is unstable. To mitigate error accumulation, recent works like **RoTTA** (Yuan et al., 2023) and **SAR** (Niu et al., 2023) introduce robustness-oriented optimization and memory banks, EATA (Niu et al., 2022) introduces sample-efficient optimization by filtering high-entropy redundancy. Furthermore, recent works have expanded TTA to handle unreliable observations in non-stationary streams (Lee and Chang, 2024) and complex multimodal tasks (Zhou et al., 2025; He et al., 2026). However, most discriminative TTA methods fundamentally rely on the "high confidence implies correctness" assumption, making them susceptible to calibration errors when decision boundaries shift (Guo et al., 2025; Marsden et al., 2023).

2.3 Probabilistic and Generative Adaptation

To address the reliability issues of discriminative TTA, recent research has pivoted towards probabilistic and generative frameworks. Probabilistic TTA methods, such as PETAL (Brahma and Rai, 2023) and Variational Neighbor-Labeling (Ambekar et al., 2024), treat pseudo-labels as latent variables, utilizing Bayesian inference to better quantify aleatoric uncertainty. **Diffusion-based TTA** methods, such as DDA (Gao et al., 2023) and Diffusion-TTA (Prabhudesai et al., 2023), utilize diffusion models to project shifted inputs back to the source manifold or provide generative feedback. While robust, these methods typically require expensive iterative denoising steps for every test sample, limiting their online applicability. Concurrently, probabilistic methods like PETAL (Brahma and Rai, 2023) use Bayesian inference to quantify uncertainty. In robotics, ADPro (Li et al., 2025a) employs manifold-constrained denoising for policy adaptation. Similarly, GIPSO (Saltori et al., 2022) leverages geometric priors for LiDAR segmentation. GD-Adapt draws inspiration from these generative and geometric insights but innovates by distilling the robustness of a generative teacher into a low-latency Student-side discriminative reader, achieving both safety and speed.

3 Methodology

We propose **GD-Adapt (Generative-Discriminative Adaptation)**, a framework designed to bridge the gap between robust generative modeling and efficient discriminative inference. The core intuition is to utilize a generative diffusion head as a robust "Teacher" to guide a lightweight regression "Student" during test-time adaptation (TTA), filtered by a manifold-aware mechanism. The overall framework is illustrated in Figure 2.

3.1 The Hybrid GD-Adapt Architecture

Our model consists of three primary components sharing a unified multimodal encoder: a Discriminative Head (Student), a Generative Head (Teacher), and a Source Manifold Estimator (SME).

Multimodal Encoder. Let $X = \{x_t, x_a, x_v\}$ represent the input modalities for text, audio, and vision. We utilize modality-specific Transformer encoders followed by a fusion Transformer to obtain a joint latent representation $z \in \mathbb{R}^d$:

$$z = \text{Encoder}(x_t, x_a, x_v) \tag{1}$$

where z serves as the conditioning variable for subsequent heads.

Generative Head (Teacher). We employ a Conditional Diffusion Model to model the probability density of the target labels y conditioned on z. Following the standard DDPM formulation, the forward process adds Gaussian noise to the label y_0 over T timesteps. The reverse process (denoising) is parameterized by a neural network $\epsilon_\theta(y_t, t, z)$ which predicts the noise component:

$$\mathcal{L}_{diff} = \mathbb{E}_{y_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}y_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t, z)\right\|^2\right] \tag{2}$$

where $\bar{\alpha}_t$ constitutes the noise schedule. This head provides robust, uncertainty-aware predictions but is computationally expensive for low-latency Student-side inference.

Discriminative Head (Student). To enable low-latency prediction, we attach a lightweight Multi-Layer Perceptron (MLP) regressor $R_\phi(z)$ to the shared encoder. This head provides deterministic predictions $\hat{y} = R_\phi(z)$ directly.

Source Manifold Estimator (SME). To detect out-of-distribution (OOD) samples that violate the source manifold assumption, we integrate a Normalizing Flow network (RealNVP-based) F_ψ. During pretraining, F_ψ learns the exact log-likelihood of source features z:

$$\log P_{src}(z) = \log P_{base}(F_\psi(z)) + \log \left|\det \frac{\partial F_\psi(z)}{\partial z}\right| \tag{3}$$

where P_{base} is a standard Gaussian distribution.

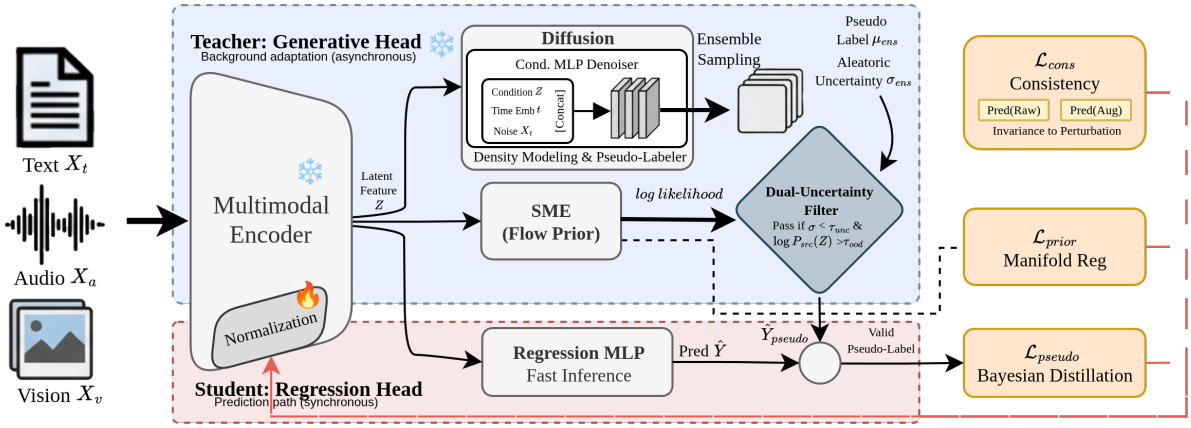

Figure 2: **Overview of the GD-Adapt Framework.** (Top) The **Generative Head (Teacher)** utilizes a Conditional Diffusion Model to learn the robust $P(y|z)$ and is used to generate uncertainty-calibrated pseudo-labels via ensemble sampling. The **Source Manifold Estimator (SME)** evaluates the log-likelihood of features to reject OOD samples. (Bottom) The **Discriminative Head (Student)** is a lightweight regression MLP optimized for fast inference. During TTA, valid samples filter through the *Dual-Uncertainty Filter* to update the Student via Bayesian Distillation, ensuring efficiency without sacrificing robustness.

3.2 Stage 1: Robustness-Oriented Pretraining

Standard pretraining often leads to overfitting source-specific correlations. We introduce Auxiliary Generative Regularization (AGR) and Stochastic Modality Perturbation (SMP) to force the encoder to learn semantic-rich, reconstructable features.

Stochastic Modality Perturbation (SMP). Specifically, we apply a hard augmentation strategy where, for a given batch, one modality is randomly masked (set to zero) and Gaussian noise is injected into the remaining modalities. This forces the model to learn robust cross-modal dependencies.

Label-Aware Supervised Contrastive Loss. To further structure the latent space for regression, we introduce a label-aware Supervised Contrastive Loss ($\mathcal{L}_{con}$). Unlike standard classification contrastive losses, we define positive pairs based on label proximity. Let z_i, z_j be normalized feature vectors and y_i, y_j be their labels. We define the set of positive pairs $\mathcal{P}(i) = \{j \mid |y_i - y_j| < \delta_{pos}\}$ and valid contrastive candidates $\mathcal{V}(i) = \{j \mid |y_i - y_j| < \delta_{pos} \vee |y_i - y_j| > \delta_{neg}\}$. This "dead zone" mechanism ($\delta_{pos} < \Delta y < \delta_{neg}$) prevents ambiguous samples from confusing the embedding space. The loss is defined as:

$$\mathcal{L}_{con} = -\sum_{i \in I} \frac{1}{|\mathcal{P}(i)|} \sum_{p \in \mathcal{P}(i)} \log \frac{e^{z_i \cdot z_p / \tau}}{\sum_{k \in \mathcal{V}(i)} e^{z_i \cdot z_k / \tau}} \quad (4)$$

where τ is the temperature parameter.

Joint Optimization. The model is trained on source data $\mathcal{D}_S$ using a composite objective:

$$\mathcal{L}_{pretrain} = \mathcal{L}_{reg} + \lambda_{gen}\mathcal{L}_{diff} + \lambda_{con}\mathcal{L}_{con} \quad (5)$$

where $\mathcal{L}_{reg}$ is the MSE loss of the discriminative head.

Flow Prior Training. After the encoder converges, we freeze the encoder and train the SME module F_ψ to maximize the likelihood of the source features, establishing a density threshold for the subsequent adaptation phase.

3.3 Stage 2: Bayesian Diffusion Distillation (BDD)

During the deployment phase, the model encounters a continuous stream of unlabeled data from a target domain $\mathcal{D}_T = \{x_t\}_{t=1}^N$. To bridge the domain gap online, we freeze the parameters of the Generative Head (Diffusion) and the SME (Flow), limiting them to act as static, calibrated experts. We adopt a parameter-efficient adaptation strategy, updating only the affine parameters of the encoder's Normalization Layers (LayerNorm) and the weights of the Student regression head (R_ϕ). This selective update strategy, inspired by Tent, allows the model to adjust to the target distribution's statistics while preserving the semantic features learned during pretraining.

3.3.1 Dual-Uncertainty Filtering (DUF)

Standard pseudo-labeling assumes that prediction confidence correlates with accuracy. However, under severe domain shifts, discriminative models often suffer from Type II errors—being "confident but wrong." To mitigate this, we propose DUF, a gating mechanism that filters samples based on two distinct types of uncertainty:

1. Aleatoric Uncertainty (Stability). This arises from inherent data ambiguity . We estimate this by querying the Diffusion Teacher K times via ancestral sampling. For an input x, we obtain a set of stochastic predictions $\{\hat{y}_1, \ldots, \hat{y}_K\}$. The ensemble mean $\mu = \frac{1}{K}\sum \hat{y}_k$ serves as the robust pseudo-label, while the standard deviation $\sigma = \text{std}(\hat{y}_k)$ quantifies the prediction instability. High σ indicates the input is ambiguous and unsuitable for supervision.

2. Epistemic Uncertainty (Manifold Adherence). This type of uncertainty arises when a target sample falls outside the source manifold learned during pre-training. We quantify it using the exact SME log-likelihood $\ell = \log P_{\text{source}}(z)$. A low likelihood indicates that the feature z has drifted away from the source support, suggesting that the Teacher's prediction may be unreliable for adaptation.

A sample is considered valid for adaptation only if it passes both the stability and manifold-adherence checks, yielding the binary mask

$$M = \mathbb{I}(\sigma < \tau_{\text{unc}}) \cdot \mathbb{I}(\ell > \tau_{\text{ood}}). \tag{6}$$

The OOD threshold τ_{ood} is calibrated from source-side validation statistics. Let $\{\ell_i^{\text{src}}\}$ denote the SME log-likelihoods computed on a held-out source validation split. We set

$$\tau_{\text{ood}} = Q_{0.05}\big(\{\ell_i^{\text{src}}\}\big),$$

where $Q_q(\cdot)$ denotes the q-th quantile; i.e., τ_{ood} is the 5th percentile of the source-validation likelihoods. In contrast, the stability threshold τ_{unc} is fixed globally during adaptation. In all experiments, we use $\tau_{\text{unc}} = 0.3$ and estimate predictive uncertainty by the standard deviation over $K = 10$ stochastic diffusion samples. Importantly, neither threshold selection nor adaptation uses target-domain labels or a target-domain validation set.

3.3.2 Distillation and Adaptation Objectives

For samples passing the DUF filter ($M = 1$), we employ a composite objective to align the efficient Student with the robust Teacher while maintaining structural constraints.

Bayesian Distillation Loss. Instead of using hard pseudo-labels, we supervise the Student regressor R_ϕ using the Teacher's ensemble mean μ. The diffusion-generated μ effectively smooths out label noise and provides a calibrated regression target:

$$\mathcal{L}_{pseudo} = M \cdot \|R_\phi(z) - \mu\|^2 \tag{7}$$

This term transfers the generative robustness to the discriminative head.

Consistency Loss. To encourage the decision boundary to pass through low-density regions (Cluster Assumption), we enforce prediction invariance under perturbation. Let z' be the feature of an augmented view of x generated via Stochastic Modality Perturbation (SMP). We minimize the divergence between the Student's predictions on raw and augmented views:

$$\mathcal{L}_{cons} = \|R_\phi(z) - R_\phi(z')\|^2 \tag{8}$$

This regularization prevents the decision boundary from crossing high-density clusters in the latent space.

Manifold Prior Loss. A common pitfall in TTA is "catastrophic forgetting," where the feature extractor drifts too far from the pre-trained state. We use the frozen SME as a regularizer to anchor the feature z within the high-density region of the source manifold:

$$\mathcal{L}_{prior} = -\log P_{source}(z) \tag{9}$$

By maximizing the likelihood of the adapted features, we ensure that the encoder continues to produce representations compatible with the frozen heads.

Online Optimization. Given a target mini-batch, the final parameters $\Theta = \{\theta_{LN}, \phi\}$ are updated by a gradient descent step on the objective

$$\mathcal{L}_{TTA} = \mathcal{L}_{pseudo} + \lambda_{cons}\mathcal{L}_{cons} + \lambda_{prior}\mathcal{L}_{prior}. \tag{10}$$

Equation (10) specifies the loss for a single mini-batch update. In our experimental protocol, this update is applied sequentially over the unlabeled target adaptation loader and repeated for multiple passes; unless otherwise stated, we use 30 epochs over the target adaptation split.

Table 1: **Comparison of performance under cross-domain shifts.** We compare GD-Adapt with Source Only and recent SOTA TTA methods including Group Contrast (Roy et al., 2023), Reliable Fusion (Yang et al., 2024), and CASP (Guo et al., 2025). ↓: lower is better, ↑: higher is better. Best results are in **bold**.

Scenario	Method	Regression		Classification / Alignment			
		MAE ↓	RMSE ↓	Acc-2 ↑	F1 ↑	Acc@0.5 ↑	Acc@1.0 ↑
MOSI → SIMS	Source Only	0.6872	0.8127	51.52	51.95	37.64	70.71
	Group Contrast	0.6722	0.7866	52.19	53.54	40.35	74.56
	Reliable Fusion	0.6867	0.8034	50.18	48.19	38.54	73.88
	CASP	0.5788	1.0617	51.21	54.01	32.85	61.02
	GD-Adapt (Ours)	**0.5673**	**0.7779**	**57.33**	**58.75**	**42.01**	**78.12**
MOSEI → SIMS	Source Only	0.6991	0.7345	56.46	57.56	41.14	72.84
	Group Contrast	0.6589	0.7869	56.01	57.73	42.54	75.50
	Reliable Fusion	0.6067	0.8034	61.17	59.29	39.94	79.47
	CASP	0.6628	0.7889	64.96	**67.05**	43.86	73.25
	GD-Adapt (Ours)	**0.5999**	**0.7014**	**66.96**	62.22	**44.61**	**88.40**
MOSI → MOSEI	Source Only	0.3291	0.4780	67.37	67.34	73.72	96.79
	Group Contrast	0.4571	0.5309	68.08	68.07	57.80	96.56
	Reliable Fusion	0.3394	0.3568	68.52	68.33	70.19	96.80
	CASP	0.3200	0.4607	69.07	68.98	71.31	97.20
	GD-Adapt (Ours)	**0.2499**	**0.3158**	**69.32**	**69.31**	**88.68**	**99.66**
SIMS → MOSI	Source Only	0.6870	0.7127	47.52	48.17	47.64	80.71
	Group Contrast	0.6161	0.7424	51.86	45.54	47.60	80.79
	Reliable Fusion	0.6176	0.7329	47.31	48.06	48.65	84.22
	CASP	0.6547	0.6832	51.63	50.06	48.87	86.38
	GD-Adapt (Ours)	**0.4838**	**0.5781**	**58.70**	**51.11**	**56.49**	**92.63**
SIMS → MOSEI	Source Only	0.5677	0.7213	47.06	48.37	47.62	70.71
	Group Contrast	0.6223	0.7239	46.23	43.18	38.43	84.34
	Reliable Fusion	0.4867	0.5234	48.63	49.96	50.12	89.30
	CASP	0.5325	0.7018	50.10	**59.60**	49.39	81.46
	GD-Adapt (Ours)	**0.3978**	**0.4727**	**50.83**	56.37	**69.59**	**97.61**

4 Experiments

4.1 Experimental Setup

Datasets and Protocols. We evaluate GD-Adapt on three standard multimodal sentiment analysis datasets: MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and SIMS (Yu et al., 2020). To rigorously assess robustness against distribution shifts, we adopt a **cross-domain** protocol. Models are pretrained on a source domain and adapted/evaluated on a different target domain without accessing target labels. We report results on five transfer scenarios: MOSI → SIMS, MOSEI → SIMS, MOSI → MOSEI, SIMS → MOSI, and SIMS → MOSEI. Note that we exclude the MOSEI → MOSI scenario because MOSEI is an extension of MOSI covering a wider range of topics, which effectively constitutes an in-domain superset evaluation rather than a challenging domain shift.

Baselines. We compare our method against Source Only and three adaptation strategies:

- **Group Contrast (GC)** (Roy et al., 2023): Originally for Image Quality Assessment, we adapt its group-wise contrastive loss to MSA. It separates high- and low-confidence samples in the latent space. *Note that GC lacks explicit mechanisms to handle multimodal heterogeneity.*
- **Reliable Fusion (RF)** (Yang et al., 2024): A multi-modal TTA method that addresses reliability bias by dynamically weighting modality contributions. *However, RF is purely discriminative and susceptible to calibration errors.*
- **CASP** (Guo et al., 2025): A state-of-the-art method utilizing context-aware self-supervised learning.

Implementation Details. We utilize a Late-Fusion Transformer architecture as the backbone

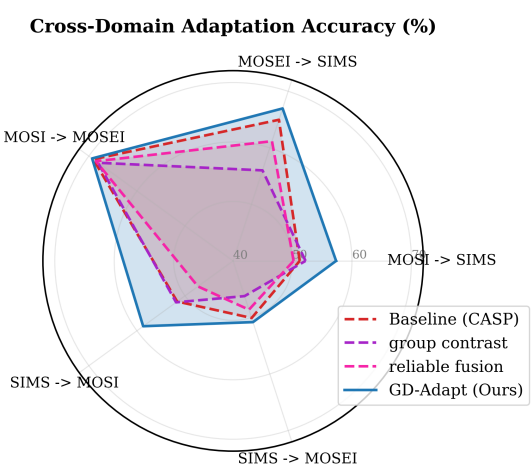

Figure 3: **Holistic Performance.** GD-Adapt (Blue) consistently encompasses Baselines across 5 scenarios.

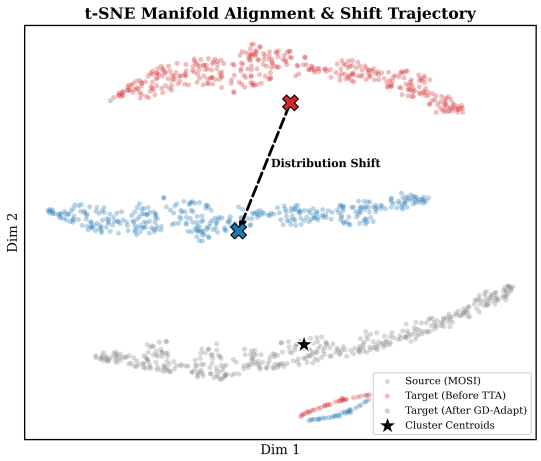

Figure 4: **t-SNE Visualization.** Left: Before TTA, severe domain shift exists. Right: GD-Adapt successfully aligns target features (Blue) to Source Manifold (Gray).

with a projection dimension of 40 and 4 attention heads. The Diffusion Teacher is pretrained with a standard schedule and employs $T = 30$ sampling steps for pseudo-label generation. During the TTA phase, we optimize the model using the **Adam** optimizer. Based on the specific domain shift, the learning rate is set to $\eta \in \{1e-3, 2e-3\}$ and the batch size is set to either 64 or 128 . Under our benchmark protocol, adaptation is performed by iterating over the unlabeled target adaptation loader for 30 epochs, applying Eq. (10) to each mini-batch. All experiments were conducted on a workstation running Ubuntu 24.04, equipped with an Intel Core Ultra Processor and a single NVIDIA RTX 5090 (32GB) GPU.

4.2 Main Results

Table 1 summarizes the performance across five cross-domain scenarios. Additionally, Fig. 3 provides a holistic visual comparison, highlighting our consistent superiority across diverse domain pairs. GD-Adapt achieves the strongest overall regression performance across all five shifts, while remaining competitive on thresholded classification metrics.

Quantitative Analysis. As shown in Table 1, GD-Adapt significantly mitigates the domain shift in two key aspects:

- **Regression Superiority:** Our method achieves the lowest MAE and RMSE across all scenarios. For instance, in SIMS $\rightarrow$ MOSI, we reduce MAE by **26.1%** compared to CASP (0.6547 $\rightarrow$ 0.4838). This confirms that the diffusion teacher successfully guides the student to the correct value manifold, correcting the "confident but wrong" offset prevalent in discriminative baselines.
- **Classification Robustness:** In challenging shifts like MOSI $\rightarrow$ SIMS, GD-Adapt improves Binary Accuracy by 5.81 percentage points over Source Only. The remarkably high Acc@1.0 scores (often $> 90\%$) further indicate that even when the binary boundary is crossed, our predictions remain highly concentrated around the ground truth.

Handling Severe Shifts. It is worth noting that performance gains are most pronounced in scenarios with large domain gaps, such as SIMS $\rightarrow$ MOSI (unscripted wild data to scripted studio data). Several baselines remain close to chance level under this severe shift, and even the strongest prior baseline remains substantially below GD-Adapt. In contrast, GD-Adapt recovers the performance to 58.70%, suggesting that our *generative* teacher effectively bridges large semantic discrepancies that purely *discriminative* methods struggle to cross.

Manifold Alignment. To visualize the adaptation process, we plot the t-SNE embeddings of the multimodal features in Fig. 4. Before adaptation (Left), the target domain distribution (SIMS) significantly deviates from the source domain (MOSI). After adaptation (Right), GD-Adapt successfully aligns the target features with the source manifold, verifying the efficacy of our Source Manifold Estimator (SME).

4.3 Inference and Adaptation Overheads

To avoid ambiguity, we distinguish latency-critical prediction from background adaptation. During deployment, predictions are served synchronously by the Student path (shared encoder + regression head), whereas Teacher-side operations, including diffusion sampling and manifold filtering, are executed asynchronously on buffered target batches and therefore do not lie on the critical prediction path.

Table 2 summarizes a compact deployment profile on the same hardware as the main experiments. The Student path remains lightweight, requiring only 0.04 ms per sample, while Teacher sampling is substantially slower (6.07 ms/sample with $T = 30$), which motivates our design of restricting the Teacher to background supervision. In addition, only 0.003M parameters are updated during test-time adaptation, and the overall memory footprint remains moderate.

We therefore interpret GD-Adapt as enabling low-latency Student-side inference with amortized background adaptation, rather than claiming that the full adaptation loop is low-latency on a per-sample basis.

Metric	Value
Student prediction (bs=1)	0.04 ms/sample
Teacher sampling (T=30, bs=1)	6.07 ms/sample
Trainable params in TTA	0.003M
Peak VRAM in TTA	4195.81 MB
System RAM in TTA	3.20 GB

Table 2: Compact deployment profile of GD-Adapt. The Student path is used for latency-critical prediction, while Teacher-based adaptation is executed asynchronously in the background.

4.4 Ablation and Sensitivity

To investigate the contribution of each component, we perform an ablation study on MOSI → SIMS. The quantitative results are listed in Table 3, and graphically illustrated in Fig. 5.

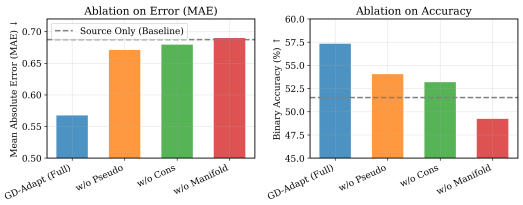

Figure 5: **Ablation Study.** Removing Manifold Prior (SME) causes the severest degradation.

Table 3: Ablation study on MOSI → SIMS. The metric values indicate the degradation when components are removed.

Variant	MAE ↓	RMSE ↓	Acc-2 ↑	Acc@1.0 ↑
Full GD-Adapt	**0.5673**	**0.7779**	**57.33**	**78.12**
w/o Pseudo-Label	0.6709	0.8030	54.05	74.40
w/o Consistency	0.6794	0.8159	53.17	73.74
w/o Manifold Prior	0.6897	0.8057	49.23	71.77

The results reveal:

- **Impact of Manifold Prior:** Removing the SME (Manifold Prior) causes the severest performance drop (MAE increases to 0.6897). This confirms that without OOD filtering, the model adapts to noisy samples ("Type II errors"), leading to catastrophic forgetting of the source distribution.
- **Impact of Consistency:** Although less critical than the Manifold Prior, consistency loss is essential for stability. As shown in Table 3, removing it increases RMSE to 0.8159. This indicates that without enforcing invariance to perturbation, the student model becomes sensitive to small input noises, resulting in unstable decision boundaries.

Hyperparameter Sensitivity. We analyze the sensitivity of the Flow Quantile Threshold τ_{ood} in Fig. 6. A low threshold admits too much noise (OOD samples), while a high threshold rejects too many valid adaptation signals. The optimal performance is observed around the 0.05 quantile, demonstrating a robust "sweet spot" for adaptation.

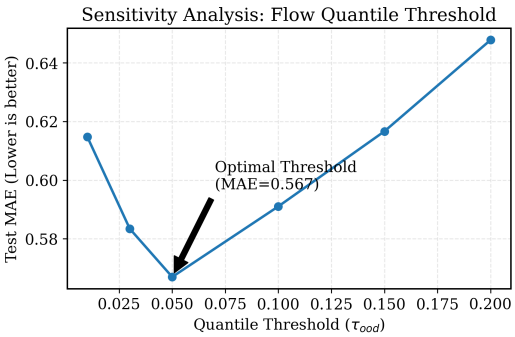

Figure 6: **Sensitivity Analysis.** MAE vs. Flow Quantile Threshold τ_{ood}. Optimal is ≈ 0.05.

5 Conclusion

In this paper, we presented GD-Adapt, a novel framework that bridges the gap between the robustness of generative models and the efficiency

of discriminative models for Test-Time Adaptation in Multimodal Sentiment Analysis. By introducing Auxiliary Generative Regularization during pretraining and employing a Source Manifold Estimator during adaptation, our method effectively mitigates the "confident but wrong" failure mode of standard TTA. Furthermore, our proposed Bayesian Diffusion Distillation strategy successfully transfers the uncertainty calibration of a diffusion teacher to a lightweight student regressor. Extensive experiments across diverse cross-domain scenarios demonstrate that GD-Adapt significantly outperforms existing baselines in both accuracy and stability while enabling low-latency student-side inference.

6 Limitations

Despite the promising performance and inference efficiency of GD-Adapt, we acknowledge several limitations that merit future investigation:

Computational Overhead During Adaptation. While our distilled Student model achieves real-time inference speed (0.04 ms/sample), the *online adaptation process* itself remains computationally demanding. The pseudo-label generation step requires invoking the Diffusion Teacher ($T = 30$ sampling steps) and the Flow network for every adaptation batch. Although this overhead does not affect the final deployed model, it potentially limits the deployment of GD-Adapt on extreme edge devices with strict power constraints during the "learning-on-the-fly" phase.

Sensitivity to Manifold Thresholds. As illustrated in our sensitivity analysis (Fig. 6), the performance of GD-Adapt relies on the appropriate selection of the out-of-distribution (OOD) threshold τ_{ood} for the Source Manifold Estimator. While setting τ_{ood} based on source validation percentiles (e.g., 5%) proves effective for the tested benchmarks, finding an optimal threshold for unknown target domains without access to a validation set remains an open challenge. An overly aggressive threshold may reject valid adaptation signals, while a loose threshold may admit noise.

Dependency on Manifold Overlap. Our method assumes a partial overlap between the source and target manifolds to trigger the adaptation. In scenarios with extreme domain shifts where the Flow model rejects the majority of incoming samples due to low likelihood, the adaptation process may stagnate (i.e., the "cold-start" problem). Future work could explore integrating few-shot retrieval mechanism to bridge disjoint manifolds.

7 Ethical Considerations

Our work focuses on improving the robustness of Multimodal Sentiment Analysis (MSA) systems under distribution shifts. We utilize publicly available academic datasets (MOSI, MOSEI, and SIMS) which contain video, audio, and text modalities. We strictly adhere to the usage licenses of these datasets and do not collect any new personally identifiable information.

However, we acknowledge that MSA technologies can carry potential risks if misused, such as in unauthorized surveillance or emotional manipulation. Although our contribution is algorithmic (improving robustness and efficiency) rather than application-specific, researchers and practitioners should exercise caution and obtain informed consent when deploying such models in real-world interactions. Additionally, while GD-Adapt reduces inference latency compared to full diffusion models, contributing to "Green AI" by lowering energy consumption during deployment, the training of diffusion models still incurs a computational carbon footprint that should be considered.

References

Sameer Ambekar, Zehao Xiao, Jiayi Shen, Xiantong Zhen, and Cees G. M. Snoek. 2024. Probabilistic Test-Time Generalization by Variational Neighbor-Labeling. *Preprint*, arXiv:2307.04033.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of LREC*, 10.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. 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)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Dhanajit Brahma and Piyush Rai. 2023. A Probabilistic Framework for Lifelong Test-Time Adaptation. *Preprint*, arXiv:2212.09713.

Fangming Feng, Sihang Cai, Zequn Xie, Yangyang Wu, and Tao Jin. 2026. Scene-aware spatiotemporal generalization: Towards robust temporal action detection

across domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 3903–3911.

Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. 2023. Back to the Source: Diffusion-Driven Adaptation to Test-Time Corruption. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11786–11796.

Zirun Guo, Tao Jin, Wenlong Xu, Wang Lin, and Yangyang Wu. 2025. Bridging the Gap for Test-Time Multimodal Sentiment Analysis. *arXiv preprint*.

Yongbo He, Zirun Guo, and Tao Jin. 2026. Decoupling stability and plasticity for multi-modal test-time adaptation. *arXiv preprint arXiv:2603.00574*.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Preprint*, arXiv:2006.11239.

Judy Hoffman, Trevor Darrell, and Kate Saenko. 2014. Continuous Manifold Based Adaptation for Evolving Visual Domains. pages 867–874.

S. Kiritchenko, X. Zhu, and S. M. Mohammad. 2014. Sentiment Analysis of Short Informal Texts. *Journal of Artificial Intelligence Research*, 50:723–762.

Jae-Hong Lee and Joon-Hyuk Chang. 2024. Stationary Latent Weight Inference for Unreliable Observations from Online Test-Time Adaptation.

Zezeng Li, Rui Yang, Ruochen Chen, ZhongXuan Luo, and Liming Chen. 2025a. ADPro: A Test-time Adaptive Diffusion Policy via Manifold-constrained Denoising and Task-aware Initialization for Robotic Manipulation. *Preprint*, arXiv:2508.06266.

Zuhe Li, Panbo Liu, Yushan Pan, Weiping Ding, Jun Yu, Haoran Chen, Weihua Liu, Yiming Luo, and Hao Wang. 2025b. Multimodal sentiment analysis based on disentangled representation learning and cross-modal-context association mining. *Neurocomputing*, 617:128940.

Jian Liang, Ran He, and Tieniu Tan. 2025. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts. *International Journal of Computer Vision*, 133(1):31–64.

Jian Liang, Dapeng Hu, and Jiashi Feng. 2021. Do We Really Need to Access the Source Data? Source Hypothesis Transfer for Unsupervised Domain Adaptation. *arXiv preprint*.

Robert A. Marsden, Mario Döbler, and Bin Yang. 2023. Universal Test-time Adaptation through Weight Ensembling, Diversity Weighting, and Prior Correction. *Preprint*, arXiv:2306.00650.

Jiana Meng, Yingchun Long, Yuhai Yu, Dandan Zhao, and Shuang Liu. 2019. Cross-Domain Text Sentiment Analysis Based on CNN_ft Method. *Information*, 10(5):162.

Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: harvesting opinions from the web. In *Proceedings of the 13th international conference on multimodal interfaces*, pages 169–176, Alicante Spain. ACM.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. 2022. Efficient Test-Time Model Adaptation without Forgetting. *Preprint*, arXiv:2204.02610.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. *Preprint*, arXiv:2302.12400.

Mihir Prabhudesai, Tsung-Wei Ke, Alexander C. Li, Deepak Pathak, and Katerina Fragkiadaki. 2023. Diffusion-TTA: Test-time Adaptation of Discriminative Models via Generative Feedback. *Preprint*, arXiv:2311.16102.

Subhadeep Roy, Shankhanil Mitra, Soma Biswas, and Rajiv Soundararajan. 2023. Test Time Adaptation for Blind Image Quality Assessment. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16696–16705. IEEE.

Cristiano Saltori, Evgeny Krivosheev, Stéphane Lathuilière, Nicu Sebe, Fabio Galasso, Giuseppe Fiameni, Elisa Ricci, and Fabio Poiesi. 2022. GIPSO: Geometrically Informed Propagation for Online Adaptation in 3D LiDAR Segmentation. *Preprint*, arXiv:2207.09763.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. 2020. Test-Time Training with Self-Supervision for Generalization under Distribution Shifts. *arXiv preprint*.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. 2021. Tent: Fully Test-time Adaptation by Entropy Minimization. *arXiv preprint*.

Hongbin Wang, Chun Ren, and Zhengtao Yu. 2025a. Multimodal sentiment analysis based on multiple attention. *Engineering Applications of Artificial Intelligence*, 140:109731.

Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. 2025b. In Search of Lost Online Test-Time Adaptation: A Survey. 133(3):1106–1139.

Zequn Xie, Boyun Zhang, Yuxiao Lin, and Tao Jin. 2026. Delving deeper: Hierarchical visual perception for robust video-text retrieval. *arXiv preprint arXiv:2601.12768*.

Mouxing Yang, Yunfan Li, Changqing Zhang, Peng Hu, and Xi Peng. 2024. Test-time adaptation against multi-modal reliability bias. In *International Conference on Representation Learning*, volume 2024, pages 55036–55055.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese Multimodal Sentiment Analysis Dataset with Fine-grained Annotation of Modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Longhui Yuan, Binhui Xie, and Shuang Li. 2023. Robust Test-Time Adaptation in Dynamic Scenarios. *Preprint*, arXiv:2303.13899.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2023. Domain Generalization: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4396–4415.

Lihua Zhou, Mao Ye, Shuaifeng Li, Nianxin Li, Jinlin Wu, Xiatian Zhu, Lei Deng, Hongbin Liu, Jiebo Luo, and Zhen Lei. 2025. Bayesian Test-time Adaptation for Object Recognition and Detection with Vision-language Models. *Preprint*, arXiv:2510.02750.