

Article

Three-Way Multimodal Learning with Severely Missing Modalities

Hanrui Wang , Yu Fang * , Xin Wang  and Fan Min 

School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China; 202321000490@stu.swpu.edu.cn (H.W.); xinwang@swpu.edu.cn (X.W.); minfan@swpu.edu.cn (F.M.)

* Correspondence: fangyu@swpu.edu.cn

Abstract

Missing modalities remain a major obstacle to the real-world deployment of multimodal learning systems, as incomplete inputs can substantially degrade model performance. Existing methods often suffer from biased imputation under high missing rates and lack uncertainty-aware, differentiated processing. Inspired by three-way decision, a framework for handling uncertainty by adding a deferment option to acceptance and rejection, we propose three-way multimodal learning with severely missing modalities (3WML-SMMs), a novel framework that introduces a three-way decision mechanism into both missing-modality imputation and feature regularization for the first time. Specifically, 3WML-SMM treats variance not merely as a descriptive measure of uncertainty, but as a decision signal for adaptive processing. Based on this idea, the framework incorporates (1) a variance-guided three-way imputation strategy with accept–delay–reject decisions to reduce unreliable reconstruction when only a limited number of complete samples are available and (2) a dimension-wise adaptive feature enhancement module that performs fine-grained regularization according to perturbation uncertainty. Experiments on the CMU Multimodal Opinion Sentiment Intensity (CMU-MOSI) and Multimodal Internet Movie Database (MM-IMDb) datasets show that 3WML-SMM consistently outperforms representative baselines, including reconstruction-based methods, complete-input multimodal methods, and missing-modality-specific methods under severe missing-modality settings, with statistically significant improvements over the multimodal learning with severely missing modality (SMIL) baseline ($p < 0.05$). These results demonstrate the effectiveness of the proposed framework, even in extreme settings where only 10% of the text modality is available.

Keywords: multimodal learning; modality missing; three-way decision; feature reconstruction; regularization; robustness



Academic Editor: Gabriele Gianini

Received: 9 March 2026

Revised: 11 April 2026

Accepted: 16 April 2026

Published: 19 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Multimodal learning has achieved remarkable success in tasks such as sentiment analysis and content understanding by integrating complementary information from diverse modalities [1,2], including images, text, and audio. Traditional multimodal approaches typically rely on a strong assumption: that all training and testing samples contain complete and aligned modalities [3]. However, this assumption is often violated in real-world applications. Due to equipment costs, privacy concerns, and difficulties in data collection, real-world datasets often suffer from severely missing modalities.

Existing research on missing modalities mainly follows two directions. One line of work uses generative models, such as autoencoders (AEs) [4] and generative adversarial networks (GANs) [5], to reconstruct missing modalities. However, under high missing rates, these methods often suffer from biased reconstruction because only a limited number of complete samples are available for effective training. The other line of work adopts meta-learning frameworks, such as SMIL [6], which is built on Bayesian meta-learning [7], to learn robust representations under missing-modality settings. Although such methods alleviate reconstruction bias to some extent, they still rely largely on uniform processing and cannot adapt reconstruction or regularization to the risk levels of individual samples or feature dimensions. Therefore, a clear research gap remains: existing methods still lack an uncertainty-aware and sample-adaptive mechanism for missing-modality reconstruction and feature regularization under severe missing conditions.

It should be noted that some existing methods are also uncertainty-aware. For example, SMIL, which is built on Bayesian meta-learning, estimates variance during reconstruction and regularization to quantify uncertainty under missing modalities. Related ideas have also appeared in probabilistic approaches such as Probabilistic Conformal Distillation (PCD) [8], which models missing-modality features as probability distributions to account for the inherent indeterminacy of mapping incomplete representations to complete ones. However, in these approaches, uncertainty is mainly used to estimate confidence or adjust the strength of a predefined operation, whereas our method uses uncertainty as an explicit decision signal to trigger qualitatively different processing paths. This limitation suggests that the key challenge is not merely how to estimate uncertainty but how to translate uncertainty into differentiated actions for cases with different risk levels.

In recent years, three-way decision theory has provided a novel perspective for handling incomplete information [9]. Formally grounded in the Trisecting–Acting–Outcome (TAO) model [10] and illustrated in Figure 1, three-way decision provides a region-based framework that partitions uncertain cases into three groups and assigns different actions accordingly. By enabling this refined form of decision-making, three-way decision has demonstrated advantages in domains such as recommender systems and anomaly detection. However, its potential to support region-specific and uncertainty-aware processing in multimodal learning, particularly for modality imputation and feature regularization, remains largely unexplored. Because it explicitly links different uncertainty levels to different actions, three-way decision provides a natural conceptual bridge from uncertainty estimation to adaptive processing in missing-modality multimodal learning.

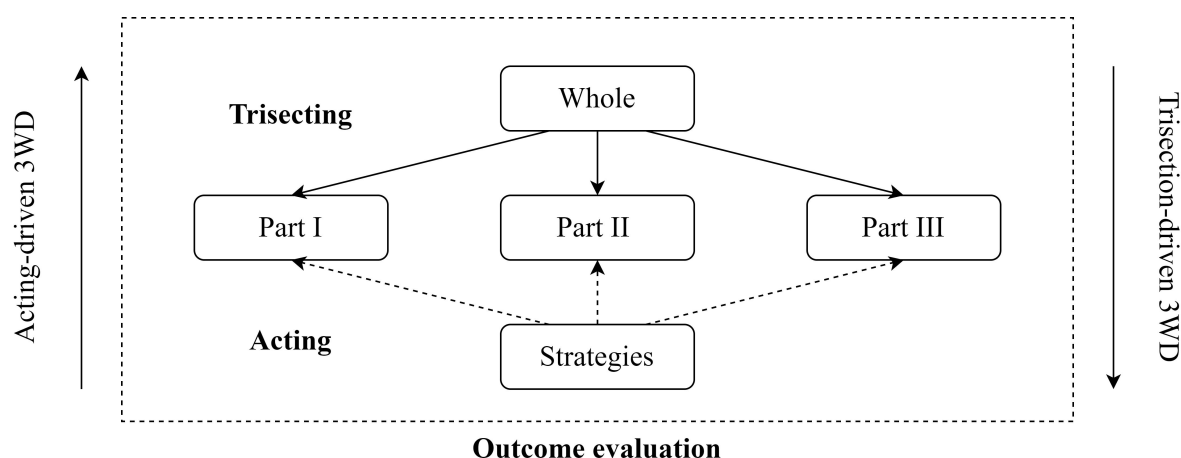


Figure 1. The conceptual framework of the three-way decision TAO model (redrawn and adapted from [10]). It illustrates the trisection-driven and acting-driven cycles that form the theoretical basis of the proposed 3WML-SMM. The function of trisectioning, denoted by solid lines, is to divide a whole

into three related and relatively independent parts. The resulting three parts are called a trisection of the whole. The function of acting, denoted by dashed lines, is to apply a set of strategies to process the three parts. By trisecting a whole and acting on the resulting trisection, it would produce an expected outcome. The function of outcome evaluation is to measure the effectiveness of the results from a combined effort of trisecting and acting as enclosed by the large dashed rectangular box in the figure.

To address this gap, this paper proposes three-way multimodal learning with severely missing modalities (3WML-SMMs), a method inspired by three-way decision theory. Motivated by this distinction, the proposed method introduces explicit uncertainty-driven decision mechanisms at both the modality and feature levels. The objective of this study is therefore to develop an uncertainty-aware and sample-adaptive framework for multimodal learning under severe missing conditions by introducing differentiated decision mechanisms into missing-modality imputation and feature regularization. Specifically, this paper makes the following contributions:

1. We introduce three-way decision theory into multimodal learning with missing modalities to develop an uncertainty-aware and sample-adaptive framework for severe missing-modality settings.
2. To address biased reconstruction under high missing rates, we design a variance-guided three-way imputation method that selectively accepts reliable reconstructions, refines ambiguous samples, and rejects highly uncertain imputations.
3. To improve representation robustness under severe missing conditions, we develop a dimension-level three-way feature enhancement strategy that applies adaptive perturbations according to feature uncertainty.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 formulates the problem and presents the proposed 3WML-SMM method. Section 4 describes the experimental setup, and Section 5 reports and discusses the experimental results. Finally, Section 6 concludes this paper and outlines future research directions.

2. Related Work

Existing approaches to multimodal learning with missing modalities can be broadly categorized into generative models, tensor completion, attention mechanisms, meta-learning, and other methods.

Early studies primarily relied on generative models [11] to reconstruct missing modalities. Autoencoders reconstruct data through an encoder–decoder architecture, generative adversarial networks generate missing data through adversarial training between a generator and a discriminator, and multimodal variational autoencoders handle multimodal data by modeling joint distributions. Nevertheless, the effectiveness of such methods depends heavily on the availability of sufficient complete samples, and their reconstruction quality degrades substantially under severe missing conditions, often introducing noise and bias.

Beyond generative approaches, tensor-completion-based methods [12] treat multimodal data as an incomplete tensor and exploit prior assumptions, such as low-rank structures, to impute missing modalities. These methods typically impose constraints on the data tensor to recover missing parts. However, they often struggle to model the complex structures and semantic relationships inherent in high-dimensional multimodal data (e.g., image and text), which limits their imputation performance.

Attention-based methods [13] improve robustness to incomplete inputs by enhancing model architectures. For instance, cross-modal attention networks can handle the absence of specific modalities through masking or adaptive weight adjustment. However, standard

attention mechanisms are typically trained under the assumption that all modalities are present, and their adaptability to extreme missing scenarios remains limited.

To reduce dependence on complete data, meta-learning frameworks [14] have been introduced into this field. SMIL employs Bayesian meta-learning to handle missing modalities during both training and testing, using a reconstruction network and a regularization network for collaborative optimization [15,16]. Similarly, some studies learn modality-invariant representations through meta-learning to enhance model generalization. However, such methods typically adopt homogeneous strategies when processing different samples and fail to make fine-grained adjustments based on sample-specific characteristics, which limits their performance under complex missing patterns.

Furthermore, knowledge distillation methods [17] transfer knowledge through teacher–student networks, and their effectiveness is constrained by the quality of the teacher network. Moreover, when modalities are severely missing, the substantial gap between the teacher and student networks can hinder successful knowledge transfer. Contrastive learning [18] approaches learn robust representations by constructing positive and negative sample pairs. Although they improve representation robustness, these methods struggle to directly recover fine-grained missing-modality information that is beneficial for downstream tasks. Graph neural networks [19] model samples and modalities as nodes in a graph, propagate information, and impute missing modality features through message passing. The effectiveness of such approaches depends heavily on the quality of graph construction, which poses challenges to their stability and scalability when data distributions are complex or missing patterns vary.

Overall, existing methods differ not only in technical form but also in the role they assign to uncertainty under missing modalities. Reconstruction-oriented methods mainly focus on recovering absent information, yet their effectiveness degrades substantially when complete samples are scarce. Representation-oriented methods improve robustness to incomplete inputs, but they usually process all samples within a shared pipeline and offer limited adaptability to the heterogeneous reliability of severely incomplete data.

Among existing approaches, uncertainty-aware meta-learning methods are most closely related to our study because they explicitly estimate uncertainty during reconstruction and regularization. In this line of research, uncertainty is generally used to estimate confidence or to adjust the strength of a predefined operation, rather than to guide qualitatively different treatments for samples or feature dimensions with different risk levels. Taken together, these studies suggest that the key distinction lies in how uncertainty is incorporated into the learning process under severe missing-modality conditions. In this context, three-way decision offers a relevant perspective for examining action-oriented uses of uncertainty in missing-modality imputation and feature regularization.

3. 3WML-SMM

3.1. Problem Statement

This study investigates multimodal learning under severe missing-modality conditions, where up to 90% of the training samples may lack one modality. Without loss of generality, this paper considers datasets containing two modalities.

Definition 1 (Multimodal Dataset). *A multimodal dataset is defined as*

$$D = \{D^f, D^m\}, \quad (1)$$

where D denotes the complete multimodal dataset consisting of two subsets: the modality-complete dataset D^f and the modality-missing dataset D^m .

Definition 2 (Modality-Complete Dataset). *The modality-complete dataset is defined as*

$$D^f = \{\mathbf{x}_i^1, \mathbf{x}_i^2, y_i\}_{i=1}^N, \quad (2)$$

where $\mathbf{x}_i^1 \in \mathbb{R}^{d_1}$ and $\mathbf{x}_i^2 \in \mathbb{R}^{d_2}$ represent the feature vectors of the first and second modalities for the i -th sample, respectively, $y_i \in \mathcal{Y}$ denotes its corresponding class label, and N is the number of modality-complete samples.

Definition 3 (Modality-Missing Dataset). *The modality-missing dataset is defined as*

$$D^m = \{\mathbf{x}_j^1, y_j\}_{j=1}^M, \quad (3)$$

where $\mathbf{x}_j^1 \in \mathbb{R}^{d_1}$ represents the available modality feature of the j -th sample, $y_j \in \mathcal{Y}$ denotes its class label, and the second modality is missing. Here, M is the number of modality-missing samples, with $M \gg N$ under severe missing conditions (e.g., $M/(N + M) \approx 90\%$).

The objective is to jointly leverage the modality-complete data D^f and the modality-missing data D^m to learn a robust multimodal model that can make accurate predictions under severe missing conditions.

A key challenge, therefore, is how to make reliable use of incomplete samples without introducing excessive uncertainty or noise into the learning process.

3.2. Motivation and Design Rationale

Under severe missing conditions, uncertainty is not only heterogeneous across samples and feature dimensions but also associated with qualitatively different risks in the subsequent learning process. Applying a uniform operation after uncertainty estimation may, therefore, be suboptimal: Reliable cases may be unnecessarily distorted, whereas unreliable cases may continue to introduce reconstruction bias or unstable perturbations into downstream representation learning. This setting calls for a mechanism that can distinguish among cases that should be preserved, further refined, or explicitly suppressed.

From this perspective, a three-way design provides a principled way to align different uncertainty levels with different intervention strategies. In missing-modality imputation, it enables the model to distinguish among directly recoverable cases, cases requiring an additional refinement stage, and cases for which their inferred modality information should not be directly propagated. In feature regularization, it likewise supports differentiated treatment of feature dimensions that require only mild perturbation, standard regularization, or stronger restraint. In this sense, three-way decision is introduced not merely as a classification heuristic but as a structural mechanism that links uncertainty estimation to action selection at both the modality and feature levels.

3.3. Model Design Process

As illustrated in Figure 2, the proposed framework consists of three components: missing-modality reconstruction, multimodal fusion, and uncertainty-aware feature regularization.

Within the Bayesian meta-learning framework, 3WML-SMM introduces two coupled three-way decision mechanisms. At the modality level, the three-way imputation router assigns each sample to the Accept, Delay, or Reject branch according to reconstruction uncertainty, producing the reconstructed modality $\hat{\mathbf{x}}^2$.

The reconstructed modality is then fused with the available modality and fed into the fusion backbone for prediction. At the feature level, the regularization net estimates dimension-wise uncertainty and applies different perturbation strategies to different feature

dimensions. In this way, 3WML-SMM performs adaptive processing at both the modality and feature levels under severe missing-modality conditions.

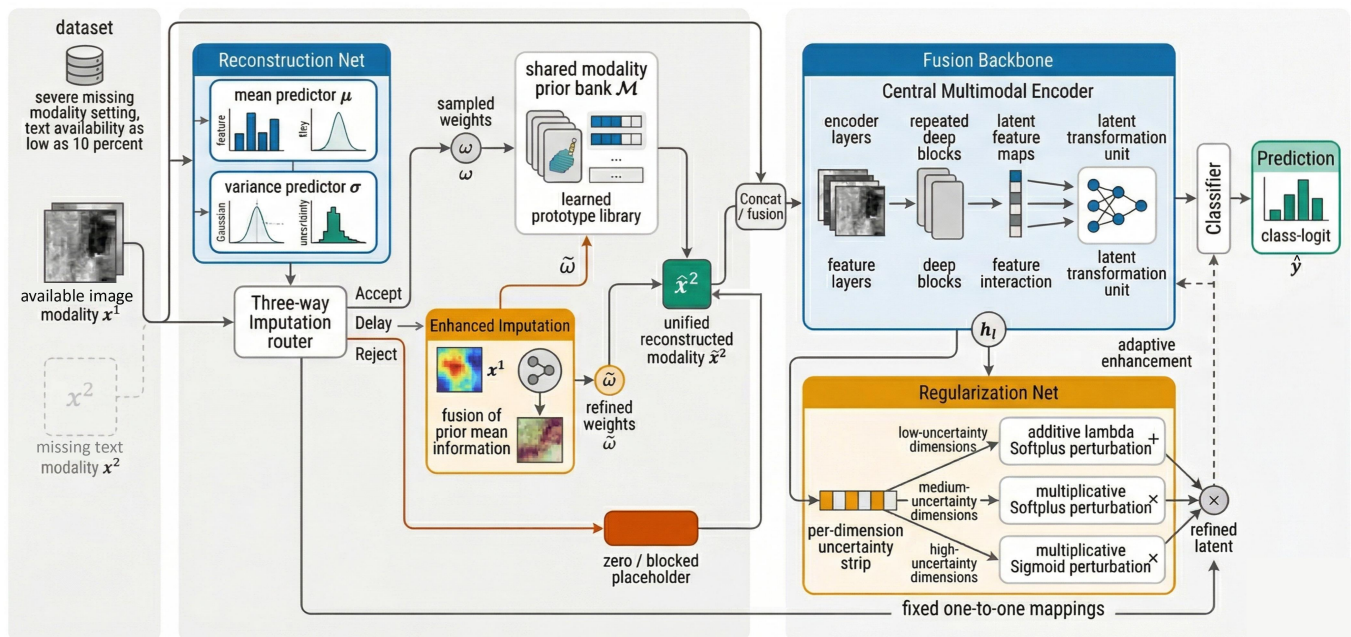


Figure 2. 3WML-SMM model architecture. In the figure, the available modality is shown in blue, the reconstructed modality in orange, the fused features in green, and the backbone network in gray. Solid arrows indicate forward data flow, while dashed arrows denote uncertainty feedback or control signals.

3.4. Three-Way Imputation

In scenarios with severe missing modalities, the key challenge is to learn a robust reconstruction model from limited complete data. The SMIL method adopts a reconstruction strategy based on prior knowledge [6], and the missing modality is reconstructed as follows:

$$\hat{x}^2 = \langle \omega, \mathcal{M} \rangle, \quad \text{where } \omega \sim \mathcal{N}(\mu, \text{diag}(\sigma)), \quad (4)$$

$\mathcal{M}$ denotes the modality prior set learned from complete data. The reconstruction network predicts the distribution parameters (μ, σ) from the available modality x^1 .

To enable uncertainty-aware strategy selection during reconstruction, this study computes the sample-level reconstruction uncertainty from the predicted variance σ :

$$u_r(\sigma) = \frac{1}{K} \mathbf{1}^T \sigma, \quad (5)$$

where $u_r(\sigma)$ denotes the sample-level reconstruction uncertainty, $\sigma \in \mathbb{R}^K$ is the variance vector of the weight distribution predicted by the reconstruction network, $K = |\mathcal{M}|$ is the number of modality priors in the modality prior set $\mathcal{M}$, and $\mathbf{1}^T$ is a row vector of all ones used to sum over all prior dimensions.

Based on this uncertainty estimate, we define a three-way imputation function $T_{\text{imp}}(\omega; \sigma, x^1, \mu)$ to determine the reconstruction action:

$$\tilde{\omega} = T_{\text{imp}}(\omega; \sigma, x^1, \mu) = \begin{cases} \omega, & \text{if } u_r(\sigma) < \beta, \\ 0, & \text{if } u_r(\sigma) > \alpha, \\ g_{\phi_g}(x^1, \mu), & \text{if } \beta < u_r(\sigma) < \alpha, \end{cases} \quad (6)$$

where α and β are decision thresholds satisfying $0 < \beta < \alpha < 1$, and g_{ϕ_g} denotes the enhanced imputation network with parameters ϕ_g . Accordingly, the uncertainty space is partitioned into three regions: acceptance ($u_r(\sigma) < \beta$), delay ($\beta < u_r(\sigma) < \alpha$), and rejection ($u_r(\sigma) > \alpha$).

Finally, the refined reconstructed features are computed as follows:

$$\hat{\mathbf{x}}^2 = \langle \tilde{\omega}, \mathcal{M} \rangle. \quad (7)$$

In this way, the reconstruction process is adaptively determined according to the estimated uncertainty of each sample.

3.5. Three-Way Regularization

In learning tasks with severe missing modalities, the robustness of feature representations is critically important. The SMIL method enhances model generalization through feature regularization, and its regularization process can be formally expressed as follows:

$$\mathbf{h}_l := \mathbf{h}_l \circ \text{Softplus}(\mathbf{r}), \quad \text{where } \mathbf{r} \sim \mathcal{N}(\boldsymbol{\mu}_r, \text{diag}(\boldsymbol{\sigma}_r)), \quad (8)$$

$\mathbf{h}_l$ denotes the latent features of the l -th layer. The regularization perturbation vector $\mathbf{r}$ is sampled from a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_r, \text{diag}(\boldsymbol{\sigma}_r))$.

The distribution parameters are computed by the regularization network f_{ϕ_r} based on the features $\mathbf{h}_{l-1}$ from the previous layer, i.e., $(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r) = f_{\phi_r}(\mathbf{h}_{l-1})$, where the mean vector $\boldsymbol{\mu}_r$ controls the base perturbation strength, and the variance vector $\boldsymbol{\sigma}_r$ characterizes the uncertainty level of each feature dimension. After mapping the perturbation to the positive domain via an activation function [20], element-wise interaction with the original features is performed to achieve adaptive feature regularization.

To enable uncertainty-aware strategy selection at the feature-dimension level, this study introduces a three-way regularization mechanism into the SMIL regularization framework. For the i -th dimension h_l^i of the l -th layer features $\mathbf{h}_l$, the uncertainty estimate σ_r^i of the perturbation is first obtained via the regularization network. Based on this, a dimension-wise three-way regularization function $T_{\text{reg}}^i(h_l^i, r^i; \sigma_r^i, \lambda)$ is defined as follows:

$$\tilde{h}_l^i = T_{\text{reg}}^i(h_l^i, r^i; \sigma_r^i, \lambda) = \begin{cases} h_l^i + \lambda \cdot \text{Softplus}(r^i), & \text{if } \sigma_r^i < \beta', \\ h_l^i \cdot \text{Sigmoid}(r^i), & \text{if } \sigma_r^i > \alpha', \\ h_l^i \cdot \text{Softplus}(r^i), & \text{if } \beta' < \sigma_r^i < \alpha', \end{cases} \quad (9)$$

where σ_r^i denotes the perturbation uncertainty of the i -th feature dimension predicted by the regularization network, α' and β' are decision thresholds satisfying $0 < \beta' < \alpha' < 1$, and λ is a decay factor ($\lambda < 1$) for weak perturbation.

For low-uncertainty dimensions with $\sigma_r^i < \beta'$, weak additive perturbation is applied to preserve feature stability. For high-uncertainty dimensions with $\sigma_r^i > \alpha'$, multiplicative perturbation with Sigmoid activation is applied to control unreliable feature variations. For moderately uncertain dimensions with $\beta' < \sigma_r^i < \alpha'$, standard multiplicative perturbation is applied to maintain balanced regularization. Finally, the regularized feature vector is expressed as follows:

$$\tilde{\mathbf{h}}_l = [\tilde{h}_l^1, \tilde{h}_l^2, \dots, \tilde{h}_l^{d_h}]^T, \quad (10)$$

where $\tilde{\mathbf{h}}_l$ denotes the regularized feature vector of the l -th layer, and d_h denotes the dimensionality of the hidden feature space. In this way, feature regularization is adaptively determined according to the uncertainty of each dimension.

3.6. Bayesian Meta-Learning

To achieve end-to-end optimization of the proposed three-way decision framework, this study extends the Bayesian meta-learning formulation adopted in SMIL. Let θ denote the parameters of the backbone network, and let $\psi = \{\phi_c, \phi_g, \phi_r\}$ collectively denote the parameters of the reconstruction network, the enhanced imputation network, and the regularization network. Let $\mathbf{z} = \{\omega, \mathbf{r}\}$ denote the set of latent variables consisting of the reconstruction weights and regularization perturbations [21]. Under this formulation, the conditional generative process of the model is written as follows:

$$p(\mathbf{Y}, \mathbf{z} | \mathbf{X}; \theta) = p(\mathbf{z}) \prod_{i=1}^N p(y_i | \mathbf{x}_i^1, \mathbf{x}_i^2, \mathbf{z}; \theta) \prod_{j=1}^M p(y_j | \mathbf{x}_j^1, \mathbf{z}; \theta), \quad (11)$$

where $\mathbf{Y}$ denotes the label set of all samples, and $\mathbf{X}$ denotes the corresponding multimodal input data.

In this framework, the reconstruction weights ω and regularization perturbations $\mathbf{r}$ are jointly modeled as latent variables $\mathbf{z} = \{\omega, \mathbf{r}\}$, where ω governs the reconstruction of missing modalities and $\mathbf{r}$ controls feature-level regularization. Accordingly, the generative process contains three components: the prior term $p(\mathbf{z})$, the likelihood term for modality-complete samples $\prod_{i=1}^N p(y_i | \mathbf{x}_i^1, \mathbf{x}_i^2, \mathbf{z}; \theta)$, and the likelihood term for modality-missing samples $\prod_{j=1}^M p(y_j | \mathbf{x}_j^1, \mathbf{z}; \theta)$.

Within this formulation, the proposed three-way imputation and regularization modules are jointly incorporated into the latent-variable inference process.

Since the true posterior $p(\mathbf{z} | \mathbf{X}, \mathbf{Y}; \theta)$ is intractable, this study adopts variational inference and uses an amortized distribution $q(\mathbf{z} | \mathbf{X}; \psi)$ to approximate it [22]. The corresponding variational lower bound is written as follows:

$$\mathcal{L}_{\text{ELBO}}(\theta, \psi) = \mathbb{E}_{q(\mathbf{z} | \mathbf{X}; \psi)} [\log p(\mathbf{Y} | \mathbf{X}, \mathbf{z}; \theta)] - \text{KL}[q(\mathbf{z} | \mathbf{X}; \psi) \| p(\mathbf{z})]. \quad (12)$$

The variational lower bound in Equation (12) consists of two terms. The first term is the expected conditional log-likelihood for label prediction, and the second term is a KL regularization term that constrains the approximate posterior toward the prior distribution.

In practice, the expectation term in Equation (12) is approximated by Monte Carlo sampling. Accordingly, the final training objective is written as follows:

$$\min_{\theta, \psi} \left(-\frac{1}{S} \sum_{s=1}^S \log p(\mathbf{Y} | \mathbf{X}, \mathbf{z}^{(s)}; \theta) + \text{KL}[q(\mathbf{z} | \mathbf{X}; \psi) \| p(\mathbf{z})] \right), \quad \mathbf{z}^{(s)} \sim q(\mathbf{z} | \mathbf{X}; \psi), \quad (13)$$

where S is the number of Monte Carlo samples. Equation (13) is written consistently at the dataset level using $(\mathbf{X}, \mathbf{Y})$, and the optimization is performed over the full training set.

During training, latent variables $\mathbf{z}^{(s)}$ are sampled from the approximate posterior $q(\mathbf{z} | \mathbf{X}; \psi)$ using the reparameterization trick, which enables gradient backpropagation through the stochastic sampling process. The negative log-likelihood term is evaluated using these sampled latent variables, while the KL divergence term is computed from the Gaussian prior and approximate posterior. The losses are aggregated across the training set to update the network parameters θ and ψ through gradient-based optimization. The thresholds $\alpha, \beta, \alpha', \beta'$ are not learned via gradient descent but are treated as hyperparameters determined by grid search on the validation set, as described in Section 4.2.

This framework jointly optimizes the network parameters through meta-learning, while the three-way decision thresholds are selected as validation-based hyperparameters.

3.7. Algorithm Design

Algorithms 1 and 2 summarize the procedural execution of the three-way imputation and three-way regularization mechanisms defined above. Their purpose is not to restate the decision rules but to clarify how the corresponding branches are instantiated during computation. In practice, the model compares real-time uncertainty estimates with predefined thresholds and assigns samples or feature dimensions to the corresponding actions. Such branch-wise execution can be implemented efficiently through parallel tensor operations.

Algorithm 1 Three-Way Imputation Algorithm

Require: Available modality feature vector $\mathbf{x}^1$, reconstruction network f_{ϕ_c} , enhanced imputation network g_{ϕ_g} , acceptance threshold β , rejection threshold α , modality prior set $\mathcal{M}$

Ensure: Reconstructed missing-modality feature $\hat{\mathbf{x}}^2$

- 1: Compute the reconstruction distribution parameters: $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = f_{\phi_c}(\mathbf{x}^1)$ by Equation (4);
 - 2: Compute the sample-level reconstruction uncertainty: $u_r(\boldsymbol{\sigma}) = \frac{1}{K} \mathbf{1}^T \boldsymbol{\sigma}$ by Equation (5);
 - 3: **if** $u_r(\boldsymbol{\sigma}) < \beta$ **then**
 - 4: Sample the reconstruction weights: $\boldsymbol{\omega} \sim \mathcal{N}(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}))$;
 - 5: Set the reconstruction weights: $\tilde{\boldsymbol{\omega}} = \boldsymbol{\omega}$;
 - 6: Perform accepted reconstruction: $\hat{\mathbf{x}}^2 = \langle \tilde{\boldsymbol{\omega}}, \mathcal{M} \rangle$ by Equation (7);
 - 7: **else if** $u_r(\boldsymbol{\sigma}) > \alpha$ **then**
 - 8: Set the reconstruction weights to zero: $\tilde{\boldsymbol{\omega}} = \mathbf{0}$;
 - 9: Perform rejected reconstruction: $\hat{\mathbf{x}}^2 = \langle \tilde{\boldsymbol{\omega}}, \mathcal{M} \rangle$ by Equation (7);
 - 10: **else**
 - 11: Obtain the refined reconstruction weights: $\tilde{\boldsymbol{\omega}} = g_{\phi_g}(\mathbf{x}^1, \boldsymbol{\mu})$ by Equation (6);
 - 12: Perform refined reconstruction: $\hat{\mathbf{x}}^2 = \langle \tilde{\boldsymbol{\omega}}, \mathcal{M} \rangle$ by Equation (7);
 - 13: **end if**
-

Algorithm 2 Three-Way Regularization Algorithm

Require: Previous-layer feature vector $\mathbf{h}_{l-1}$, current hidden feature vector $\mathbf{h}_l$, regularization network f_{ϕ_r} , acceptance threshold β' , rejection threshold α' , decay factor λ

Ensure: Regularized feature vector $\tilde{\mathbf{h}}_l$

- 1: Compute the perturbation distribution parameters via the regularization network: $(\boldsymbol{\mu}_r, \boldsymbol{\sigma}_r) = f_{\phi_r}(\mathbf{h}_{l-1})$ by Equation (8);
 - 2: Initialize the regularized feature vector: $\tilde{\mathbf{h}}_l = \mathbf{h}_l$;
 - 3: **for** each feature dimension j of $\mathbf{h}_l$ **do**
 - 4: **if** $\sigma_r^j < \beta'$ **then**
 - 5: Sample the perturbation: $r^j \sim \mathcal{N}(\mu_r^j, \sigma_r^j)$;
 - 6: Apply additive perturbation: $\tilde{h}_l^j = h_l^j + \lambda \cdot \text{Softplus}(r^j)$ by Equation (9);
 - 7: **else if** $\sigma_r^j > \alpha'$ **then**
 - 8: Sample the perturbation: $r^j \sim \mathcal{N}(\mu_r^j, \sigma_r^j)$;
 - 9: Apply multiplicative perturbation: $\tilde{h}_l^j = h_l^j \cdot \text{Sigmoid}(r^j)$ by Equation (9);
 - 10: **else**
 - 11: Sample the perturbation: $r^j \sim \mathcal{N}(\mu_r^j, \sigma_r^j)$;
 - 12: Apply multiplicative perturbation: $\tilde{h}_l^j = h_l^j \cdot \text{Softplus}(r^j)$ by Equation (9);
 - 13: **end if**
 - 14: **end for**
-

4. Experiments

4.1. Experimental Dataset

The CMU-MOSI dataset consists of 2199 opinion video clips collected from YouTube movie reviews. Each clip contains three modalities: image, text, and audio. The image modality captures visual gestures, the text modality contains transcribed speech, and the

audio modality consists of raw acoustic signals. Following Liu et al. [23], we extract features from each modality separately. The dataset is officially divided into training, validation, and test sets, with detailed statistics provided in Table 1. In this study, only the image and text modalities are used.

Table 1. Dataset split statistics for CMU-MOSI and MM-IMDb.

Dataset	Split	Number of Samples	Percentage of Total
CMU-MOSI	Training Set	1284	58.4%
	Validation Set	229	10.4%
	Test Set	686	31.2%
	Total	2199	100%
MM-IMDb	Training Set	15,552	59.9%
	Validation Set	2608	10.1%
	Test Set	7799	30.0%
	Total	25,959	100%

The MM-IMDb dataset contains two modalities: image and text. This study uses the dataset for movie genre prediction, a task that can be performed using either the image or text modality alone. Since a movie may belong to multiple genres simultaneously, this task is inherently a multi-label classification problem. The dataset comprises 25,959 movie samples covering 23 categories. The official training/validation/test split follows the scheme established by Vielzeuf et al. [24], with detailed statistics provided in Table 1.

4.2. Experimental Settings

We evaluate the proposed method on both the CMU-MOSI and MM-IMDb datasets using the parameter settings listed in Table 2. In all experiments, the image modality is assumed to be always complete, and missingness is introduced only in the text modality. This setting is adopted to provide a controlled severe-missing scenario in which the visual modality serves as a stable reference, while the model is evaluated on its ability to cope with limited text information. The text-modality completeness rate is defined as $\eta = N_{\text{text}} / N_{\text{total}}$, where N_{text} denotes the number of samples with available text modality and N_{total} is the total number of samples. A smaller η indicates more severe text-modality missingness. In this study, two settings are considered, namely, $\eta \in \{0.1, 0.2\}$, corresponding to 10% and 20% available text modalities, respectively. Equivalently, these correspond to 90% and 80% text-missing rates. For each setting, samples with available text modality are randomly selected, while the remaining samples are treated as text-missing. This missingness process is implemented through random subsampling rather than replication of existing data points. Accordingly, the modality combination for each setting is represented as 100% image + $\eta\%$ text.

Table 2. Parameter settings.

CMU-MOSI Dataset		MM-IMDb Dataset	
Parameter	Value	Parameter	Value
Batch Size	32	Batch Size	128
Optimizer	Adam	Optimizer	Adam
Inner Loop LR	1×10^{-4}	Inner Loop LR	1×10^{-4}
Outer Loop LR	1×10^{-4}	Outer Loop LR	1×10^{-3}
Iterations	5000	Iterations	10,000

For the CMU-MOSI dataset, image and text modality features are obtained following the approach of Liu et al. [23]. The image modality is embedded through three fully connected layers with a dimension of 16, while the text modality features are extracted using a single-layer LSTM [25]. The two modality features are then concatenated and fed into a fully connected layer for classification. Training is performed using the Adam optimizer [26].

For the MM-IMDb dataset, image and text modality features are extracted using the model proposed by Arevalo et al. [27]. Features from each modality are aligned in dimension through fully connected layers before fusion and are finally fed into a fully connected layer for multi-label classification. To address label imbalance, following the approach of Vielzeuf et al., a weight coefficient of 2.0 is assigned to positive samples to balance precision and recall.

Imputation Decision Thresholds: The threshold pair (α, β) is set to (0.7, 0.3). Both thresholds are constrained to the interval (0, 1) with $\beta < \alpha$, ensuring that the acceptance, delay, and rejection regions are properly nested. The imputation thresholds are tuned on the validation set by grid search over candidate values $\alpha \in \{0.5, 0.6, 0.7, 0.8\}$ and $\beta \in \{0.2, 0.3, 0.4\}$ under the constraint $0 < \beta < \alpha < 1$. The final threshold pair is selected according to the best validation performance under the corresponding evaluation metric of each dataset.

Regularization Parameters: The threshold pair (α', β') is set to (0.8, 0.2), and the decay factor is set to $\lambda = 0.5$. Similarly, the regularization thresholds are determined by validation-based grid search over predefined candidate values under the constraint $0 < \beta' < \alpha' < 1$. The final configuration is selected according to the best validation performance under the corresponding evaluation metric of each dataset so as to provide sufficient robustness for high-uncertainty features while preserving the stability of low-uncertainty features.

4.3. Baselines

Lower Bound: This baseline model is trained using only a single modality (e.g., 100% image or 100% text). Its performance serves as the lower-bound benchmark for evaluating the proposed method.

Upper Bound: This baseline model is trained using all modalities (i.e., 100% image and 100% text). Its performance is regarded as the upper-bound benchmark in this study.

Multimodal Variational Autoencoder (MVAE) [28]: MVAE is a model proposed for multimodal generative tasks. In this study, we adopt the widely used linear evaluation protocol to adapt it for classification tasks. Specifically, MVAE is first pre-trained using all modalities; then, with its model parameters frozen, the latent representations generated by its encoder are used to train a randomly initialized linear classifier.

AE [4]: AE is an autoencoder-based model used for modality reconstruction. In our experiments, AE is employed as a preprocessing method to alleviate severe missing modalities. Specifically, a subset of modality-complete samples is first selected from the original dataset. One modality is then masked during training, and the trained AE is used to reconstruct the missing modality for incomplete samples.

GAN [5]: GAN is a deep generative model consisting of a generator and a discriminator. In this study, we follow the same preprocessing procedure as that used for AE and employ GAN to address the missing-modality problem in the same manner.

Multimodal Transformer (MulT) [29]: MulT is a representative Transformer-based multimodal fusion model that employs cross-modal attention mechanisms to directly model interactions between modalities.

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (MISA) [30]: This is a method that explicitly disentangles representations into modality-invariant and modality-specific components to learn more robust representations.

Self-Supervised Multi-task Multimodal Sentiment Analysis Network (Self-MM) [31]: Self-MM is a self-supervised framework that enhances robustness to incomplete modalities by generating labels for missing modalities.

Transformer-Based Feature Reconstruction Network (TFR-Net) [32]: TFR-Net is a network specifically designed to handle missing modalities during testing by recovering missing information through feature reconstruction.

Contrastive Learning and Multi-Layer Fusion (CLMLF) [33]: CLMLF is a contrastive learning-based multimodal framework that learns robust representations under missing modalities by constructing a cross-modal contrastive loss.

CIF-Aware Missing Modality Imagination Network (CIF-MMIN) [34]: CIF-MMIN is a method that acquires modality-invariant features for multimodal sentiment recognition under missing modalities based on contrastive learning.

For fairness, all baseline methods are evaluated under the same data splits, missing-modality protocol, and evaluation metrics as the proposed method. Specifically, the image modality is always kept complete, while the text modality is made severely missing according to the tested completeness rates. Reconstruction-based baselines, such as AE and GAN, are first trained on the modality-complete subset to recover the missing text modality, after which the imputed samples are used for downstream multimodal classification under the same classifier setting. For MVAE, we follow the standard linear evaluation protocol: The model is first pre-trained using complete multimodal data, and then, its encoder is frozen while a linear classifier is trained on the resulting latent representations.

For baselines originally designed for complete multimodal inputs, such as MulT and MISA, training and testing are conducted using only the modalities available under each missing setting, without introducing additional imputation beyond the observed inputs. For methods explicitly designed for missing-modality learning, including Self-MM, TFR-Net, CLMLF, CIF-MMIN, and SMIL, we follow their original implementation principles while re-training them under the same missing-text masks, data partitions, and evaluation criteria used in our experiments. Therefore, all methods are compared under a unified and controlled experimental protocol.

4.4. Evaluation Metrics

To comprehensively evaluate the performance of the proposed 3WML-SMM method under severe missing-modality conditions, two types of evaluation metrics are used according to the task characteristics of the two datasets.

For the binary sentiment prediction task on the CMU-MOSI dataset, we use accuracy and F1 scores as evaluation metrics. Accuracy measures the proportion of correctly predicted samples among all samples, while the F1 score is the harmonic mean of precision and recall, providing a more comprehensive assessment of classification performance on imbalanced data. The formulas for these metrics are as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \quad (14)$$

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \quad (15)$$

where TP (True Positive) denotes the number of true positive samples, TN (True Negative) denotes the number of true negative samples, FP (False Positive) denotes the number of false positive samples, and FN (False Negative) denotes the number of false negative samples.

For the multi-label movie genre classification task on the MM-IMDb dataset, we use F1-Samples and F1-Micro as evaluation metrics. F1-Samples is computed at the sample level, where the F1 score is calculated for each sample and then averaged over all samples. F1-Micro computes the F1 score by globally accumulating TP, FP, and FN across all samples. These two metrics provide an effective way to evaluate the overall performance of the model on multi-label classification tasks. The formulas for these metrics are as follows:

$$F1-Samples = \frac{1}{N} \sum_{i=1}^N \frac{2 \times |Y_i \cap \hat{Y}_i|}{|Y_i| + |\hat{Y}_i|}, \quad (16)$$

$$F1-Micro = \frac{2 \times \sum_{i=1}^N |Y_i \cap \hat{Y}_i|}{\sum_{i=1}^N |Y_i| + \sum_{i=1}^N |\hat{Y}_i|}, \quad (17)$$

where N denotes the total number of samples, Y_i denotes the true label set of the i -th sample, and $\hat{Y}_i$ denotes the predicted label set of the i -th sample.

5. Results and Analysis

5.1. Comparison with the Baseline Model

To comprehensively validate the effectiveness of the proposed method (3WML-SMM) under severe missing-modality conditions, we conduct systematic experiments on the CMU-MOSI and MM-IMDb datasets. As shown in Tables 3 and 4, the proposed method is evaluated from multiple perspectives appropriate to each dataset, including accuracy and F1 score on CMU-MOSI, as well as F1-Samples and F1-Micro on MM-IMDb. Across these evaluation metrics, 3WML-SMM achieves the best performance under all tested severe missing-modality settings.

Table 3. Binary classification performance on the CMU-MOSI dataset with different text modality proportions.

Model	Accuracy (%)			F1 Score		
	10%	20%	100%	10%	20%	100%
Lower-Bound	-	-	44.8	-	-	27.7
Upper-Bound	-	-	71.0	-	-	70.5
MVAE [28]	-	-	58.5	-	-	58.1
AE [4]	56.4	60.4	-	54.4	59.0	-
GAN [5]	56.5	60.6	-	54.6	59.1	-
MuT [29]	57.5	61.0	81.5	56.8	60.5	80.2
MISA [30]	57.0	60.5	80.0	56.5	60.0	79.1
Self-MM [31]	58.2	61.8	80.8	57.7	61.3	79.5
TFR-Net [32]	58.5	62.0	79.0	58.0	61.5	77.8
CLMLF [33]	59.0	62.3	84.2	58.5	61.8	83.7
CIF-MMIN [34]	59.8	62.7	86.8	59.2	62.3	85.6
SMIL [6]	60.7	63.3	-	58.0	62.5	-
3WML-SMM (Ours)	62.5	64.6	-	60.5	64.1	-

‘-’ indicates that this setting is not applicable or does not require testing. Bold indicates the best performance in each column.

Several observations can be drawn from Tables 3 and 4. Reconstruction-based baselines such as AE and GAN generally show lower performance than the stronger multimodal and missing-modality-specific baselines under severe missing settings. Methods based on strong cross-modal interaction, such as MuT and MISA, also tend to perform less favorably when one modality is highly incomplete.

The comparison with SMIL further suggests that uncertainty estimation alone may be insufficient to ensure robust learning under severe missing conditions. Although SMIL is

the strongest comparative baseline in our experiments, 3WML-SMM consistently achieves better results across the tested severe missing settings. These results suggest that, beyond estimating uncertainty, the way uncertain information is handled during learning is also important under severe missing conditions.

Table 4. Multi-label classification performance on the MM-IMDb dataset with different text modality proportions.

Model	F1-Samples			F1-Micro		
	10%	20%	100%	10%	20%	100%
Lower-Bound	-	-	47.6	-	-	48.2
Upper-Bound	-	-	61.7	-	-	62.0
MVAE [28]	-	-	48.4	-	-	48.6
AE [4]	44.5	50.9	-	44.8	50.7	-
GAN [5]	45.0	51.1	-	44.6	51.0	-
MuT [29]	46.8	52.5	63.5	47.0	52.8	63.8
MISA [30]	46.2	51.9	62.8	46.5	52.1	63.0
Self-MM [31]	47.5	53.2	63.9	47.8	53.5	64.1
TFR-Net [32]	47.8	53.0	61.5	48.0	53.3	61.7
CLMLF [33]	48.4	53.4	64.7	48.5	53.6	65.2
CIF-MMIN [34]	48.8	53.7	68.2	48.7	54.2	68.5
SMIL [6]	49.2	54.1	-	49.5	54.6	-
3WML-SMM (Ours)	50.7	55.4	-	50.8	56.2	-

‘-’ indicates that this setting is not applicable or does not require testing. Bold indicates the best performance in each column.

5.2. Statistical Significance Analysis

To further assess the statistical reliability of the observed performance gains, we conduct a paired t -test comparing 3WML-SMM with SMIL. SMIL is chosen as the statistical comparator because it is the most directly related uncertainty-aware baseline and also one of the strongest comparative baselines under the severe missing-modality settings considered in this study. All experiments are repeated 10 times using different random seeds. For each setting (dataset, missing rate, and evaluation metric), the paired t -test is performed based on the 10 paired performance differences between 3WML-SMM and SMIL. The t -statistic is computed as follows:

$$t = \frac{\bar{d}}{\sigma_d / \sqrt{n}}, \quad (18)$$

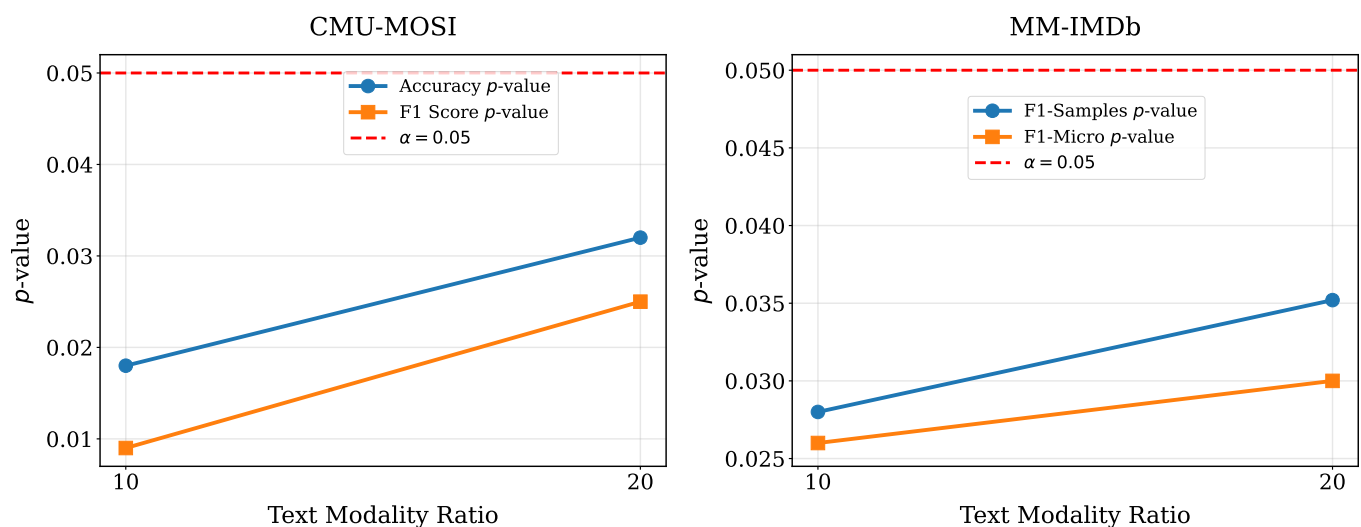
where $\bar{d}$ denotes the mean of the paired performance differences between 3WML-SMM and SMIL across repeated runs, σ_d denotes the standard deviation of these paired differences, and n denotes the number of runs. The corresponding p -value is derived from the t -distribution with $n - 1$ degrees of freedom.

As shown in Table 5 and Figure 3, the paired t -test results indicate that the performance improvements of 3WML-SMM over SMIL remain statistically significant ($p < 0.05$) in all evaluated cases.

Overall, the statistical test results support the reliability of the observed performance gains of 3WML-SMM over SMIL under severe missing conditions. These results provide complementary statistical evidence for the empirical comparisons reported in Section 5.1. The specific effects of the proposed mechanisms are further examined in the subsequent ablation and visualization analyses.

Table 5. Exact p -values of the paired t -tests for the comparison between 3WML-SMM and SMIL across all evaluated settings.

Dataset	Text Availability	Metric	p -Value
CMU-MOSI	10%	Accuracy (%)	0.018
		F1 Score	0.009
	20%	Accuracy (%)	0.032
		F1 Score	0.025
MM-IMDb	10%	F1-Samples	0.028
		F1-Micro	0.026
	20%	F1-Samples	0.035
		F1-Micro	0.030

**Figure 3.** p -values of the paired t -test for the performance improvement of 3WML-SMM over SMIL on the CMU-MOSI and MM-IMDb datasets under different missing-modality settings.

5.3. Visualization Analysis of Feature Space

This study conducts a visualization analysis of the multimodal embeddings learned by the model on the MM-IMDb dataset. Figure 4 presents a t-distributed stochastic neighbor embedding (t-SNE) visualization of the learned embedding spaces under different text-modality availability ratios. Three movie genres with substantial visual and semantic differences are selected as examples, where the red, green, and blue points denote ‘Sport’, ‘Film-Noir’, and ‘Western’, respectively.

As shown in Figure 4, in the embedding space of the image-only lower-bound baseline (Figure 4a), samples from different categories are highly mixed and exhibit limited discriminability. In contrast, in the embedding space learned by 3WML-SMM with only 10% text modality available (Figure 4b), intra-class aggregation begins to improve, resulting in clearer inter-class boundaries. This improvement becomes more pronounced when text-modality availability increases to 20% (Figure 4c), where the feature distributions exhibit substantially improved intra-class compactness and inter-class separation.

Despite the high missing rates of the text modality (90% and 80%, respectively), 3WML-SMM effectively filters and enhances the available information, thereby learning more discriminative feature representations. Notably, the visualization results of 3WML-SMM with 20% text (Figure 4c) move toward the clearer intra-class clustering and inter-class separation exhibited by the upper-bound baseline (Figure 4d), which uses all complete modalities.

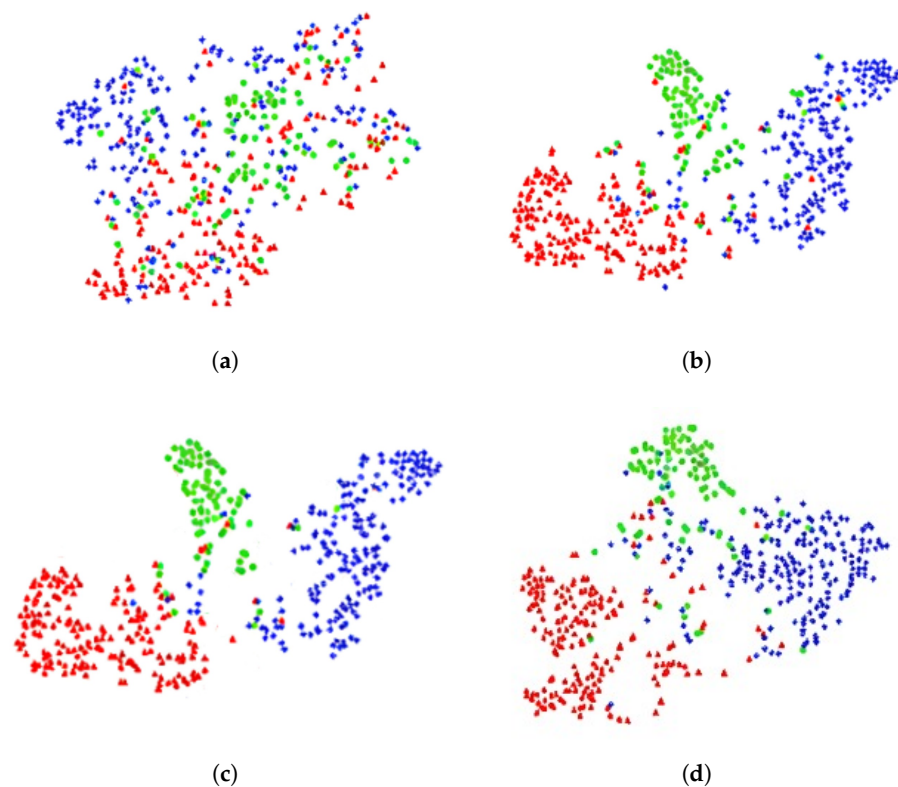


Figure 4. t-SNE visualization of feature embeddings on the MM-IMDb dataset. Different colors represent different movie genres: red for Sport, green for Film-Noir, blue for Western. Each point shape corresponds to a single data sample. (a) Lower-Bound: 100% Image. (b) Ours: 100% Image + 10% Text. (c) Ours: 100% Image + 20% Text. (d) Upper-Bound: 100% Image + 100% Text.

5.4. Ablation Study

To examine the contribution of each core component in the proposed three-way decision framework, we conduct systematic ablation experiments on the MM-IMDb dataset. As shown in Table 6, removing or simplifying key model designs leads to consistent performance decreases across all reported metrics, suggesting that each module contributes positively under severe missing conditions.

Table 6. Ablation study on the MM-IMDb dataset.

Model	F1-Samples		F1-Micro	
	10%	20%	10%	20%
w/o Both	0.492	0.541	0.495	0.546
w/ Imputation Decider	0.495	0.543	0.497	0.548
w/ Regularization Decider	0.498	0.547	0.501	0.552
w/o Delay	0.502	0.548	0.503	0.554
3WML-SMM (Full)	0.505	0.552	0.508	0.558

‘Both’ refers to the imputation decider and the regularization decider. Bold indicates the best performance in each column.

The ablation results show that the full model consistently outperforms all simplified variants across the reported metrics, suggesting that each proposed component contributes to the overall effectiveness of the full model under severe missing conditions. In particular, removing either the three-way imputation mechanism or the three-way regularization mechanism leads to lower performance than that of the full model. In addition, the performance gap between the full model and 3WML-SMM without delay suggests that the delay branch may provide useful additional flexibility beyond a binary accept–reject strategy.

To further examine the role of the delay mechanism, Table 7 reports several runtime statistics of the full 3WML-SMM model on the MM-IMDb dataset with 20% text available.

Table 7. Quantitative analysis of the ‘Delay’ decision on the MM-IMDb dataset with 20% text available.

Analysis Category	Delay Decision Statistics	Result
Decision Distribution	Delayed Region Sample Ratio	35.2%
	Accepted Region Sample Ratio	49.7%
	Rejected Region Sample Ratio	15.1%
Enhancement Effect	Samples with Significantly Improved Confidence *	82.4%
	Average Confidence Improvement	+44.7%
Performance Contribution	Relative Contribution to ‘w/o Delay’ Performance Gain +	~41.3%

‘*’ indicates a delayed sample with relative confidence improvement > 30%. ‘+’ indicates the estimated contribution of the delay mechanism to the performance gain relative to the variant without delay.

Significant Confidence Improvement *: For each delayed sample, reconstruction confidence is defined as the inverse of the sample-level reconstruction uncertainty. Let σ denote the uncertainty vector before enhanced imputation, and let σ' denote the refined uncertainty vector additionally estimated for delayed samples by the enhanced imputation network. The confidence values before and after enhanced imputation are then defined as follows:

$$C_{\text{before}} = \frac{1}{u_r(\sigma)}, \quad C_{\text{after}} = \frac{1}{u_r(\sigma')}. \quad (19)$$

A delayed sample is considered significantly improved if

$$\frac{C_{\text{after}} - C_{\text{before}}}{C_{\text{before}}} > 0.3. \quad (20)$$

The reported percentage represents the proportion of delayed samples that satisfy this condition, and the average confidence improvement is computed over all delayed samples.

Performance Gain Contribution +: This value is estimated by comparing the performance difference between the full 3WML-SMM model and the ‘w/o Delay’ variant under the same dataset setting and evaluation metric. Specifically, let P_{full} denote the performance of the full model, and let $P_{\text{w/o Delay}}$ denote the performance of the variant without the delay mechanism. The relative contribution of the delay mechanism is then computed as follows:

$$\text{Contribution} = \frac{P_{\text{full}} - P_{\text{w/o Delay}}}{P_{\text{full}}}. \quad (21)$$

This value reflects the proportion of the total performance gain attributable to the delayed-sample processing branch rather than serving as an independent evaluation metric.

Table 7 shows that 35.2% of samples are assigned to the delay region, indicating that a non-negligible proportion of samples fall within an intermediate uncertainty range under severe missingness. Among these delayed samples, 82.4% achieve significant confidence improvement after enhanced imputation, with an average increase of 44.7%. In addition, the estimated contribution of the delay branch relative to the ‘w/o Delay’ variant is approximately 41.3%, suggesting that the delay mechanism is associated with a meaningful part of the overall performance gain of the full model. Since the delay mechanism is applied only to samples routed to the delay region, it introduces additional computation selectively rather than uniformly across all samples.

As illustrated in Figure 5, the three regions exhibit distinct spatial patterns that are broadly consistent with their functional roles. Accept-region samples form a relatively

dense cluster, delay-region samples occupy an intermediate transitional zone, and reject-region samples are distributed farther from the core. This spatial pattern is consistent with the quantitative statistics in Table 7.

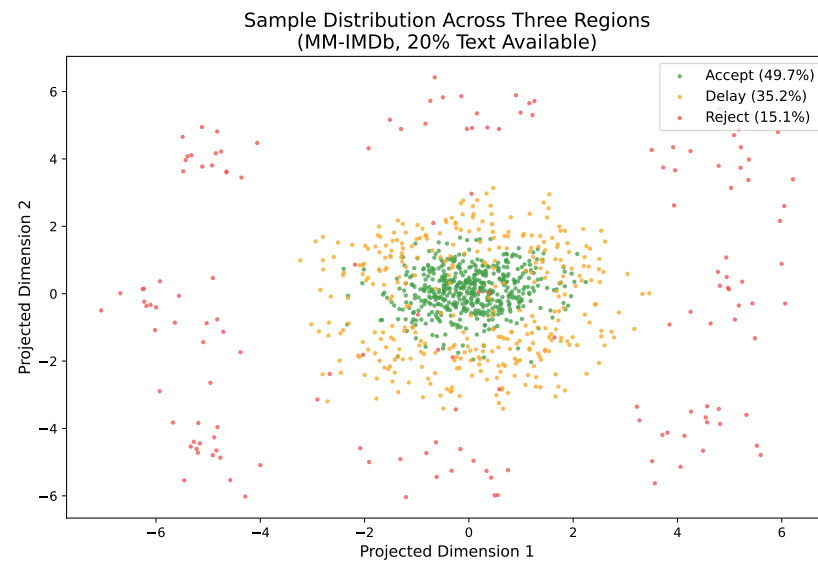


Figure 5. Sample distribution across the accept, delay, and reject regions of the three-way imputation stage on MM-IMDb with 20% text available.

Overall, the ablation results indicate that the three-way imputation decoder, the three-way regularization decoder, and the delay mechanism each contribute positively under severe missing conditions, while their combination yields the strongest overall performance among the tested variants.

5.5. Sensitivity Analysis of Imputation Thresholds

The selection of the acceptance threshold β and rejection threshold α in the three-way imputation mechanism may affect model performance. To investigate the sensitivity of 3WML-SMM to these thresholds, we conduct a grid search on the CMU-MOSI dataset under severe missing conditions ($\eta = 0.1$, i.e., 10% text available), with $\alpha \in \{0.5, 0.6, 0.7, 0.8\}$ and $\beta \in \{0.2, 0.3, 0.4\}$. The results are presented in Figure 6.

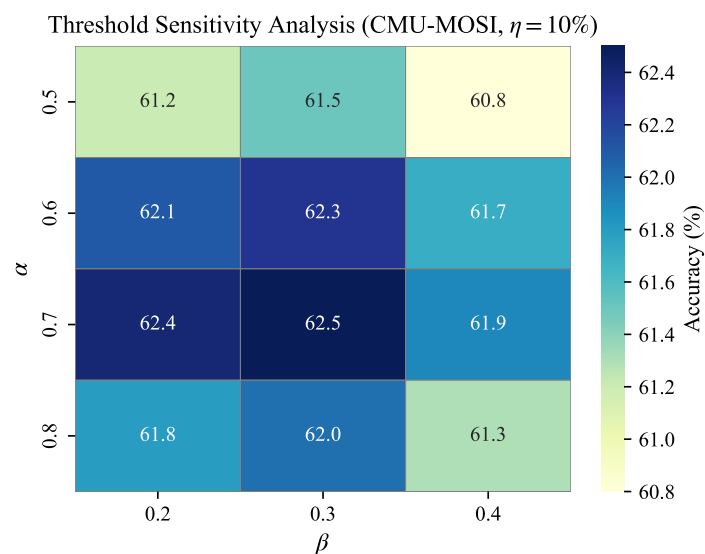


Figure 6. Threshold sensitivity analysis on CMU-MOSI with $\eta = 0.1$ (10% text available). Each cell shows the classification accuracy (%) for a given (α, β) pair; darker color indicates higher accuracy.

Figure 6 shows that the classification accuracy first increases and then decreases as the thresholds move away from the optimal setting, with the best result achieved at $(\beta = 0.3, \alpha = 0.7)$. This pattern indicates that three-way imputation requires a balance between preserving reliable samples and preventing uncertain ones from exerting excessive influence on subsequent learning. Around the optimal setting, the model remains relatively stable, maintaining accuracy above 62.0% for $\beta \in [0.2, 0.3]$ and $\alpha \in [0.6, 0.7]$. Moreover, performance varies more noticeably with β than with α , suggesting that the acceptance threshold plays a more critical role than the rejection threshold under severe missingness.

6. Conclusions

This paper investigates multimodal learning under severe missing-modality conditions and proposes 3WML-SMM, a framework that introduces three-way decision into both missing-modality imputation and feature regularization. Specifically, the proposed method uses variance-guided three-way imputation to differentiate among samples with different levels of reconstruction uncertainty and adopts dimension-level three-way regularization to perform uncertainty-aware feature enhancement.

The experimental results of the CMU-MOSI and MM-IMDb datasets show that 3WML-SMM achieves competitive performance under the evaluated severe missing-modality settings. The ablation and quantitative analyses are broadly consistent with the intended roles of the proposed modules, especially the usefulness of the delay mechanism for handling uncertain but potentially recoverable samples.

Despite these encouraging results, the current study remains limited to two benchmark datasets and a specific missing-modality setting in which the image modality is complete while the text modality is severely missing. In addition, the decision thresholds are selected as validation-based hyperparameters rather than being learned adaptively during training. Future work will therefore evaluate the proposed framework under more diverse missing patterns and more complex multimodal settings while also exploring adaptive threshold-learning strategies.

Author Contributions: Conceptualization, H.W. and Y.F.; methodology, H.W.; software, H.W.; validation, H.W., Y.F., X.W. and F.M.; formal analysis, H.W.; investigation, H.W.; resources, Y.F.; data curation, H.W.; writing—original draft preparation, H.W.; writing—review and editing, Y.F., X.W. and F.M.; visualization, H.W.; supervision, Y.F.; project administration, Y.F.; funding acquisition, Y.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant numbers 62176221 and 62276215.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are derived from publicly available datasets. The CMU-MOSI dataset is available on Kaggle at <https://www.kaggle.com/datasets/mathurinache/cmu-mosi> (accessed on 4 March 2025). The MM-IMDb dataset used in this study is a version maintained in the repository at <https://github.com/johnarevalo/gmu-mmimdb> (accessed on 4 March 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Das, R.; Singh, T.D. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Comput. Surv.* **2023**, *55*, 1–38. [\[CrossRef\]](#)

2. Wu, R.; Wang, H.; Chen, H.T.; Carneiro, G. Deep multimodal learning with missing modality: A survey. *Trans. Mach. Learn. Res.* **2026**. Available online: <https://openreview.net/forum?id=tc7RFcx4hT> (accessed on 15 April 2026).
3. Baltrušaitis, T.; Ahuja, C.; Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 423–443. [[CrossRef](#)] [[PubMed](#)]
4. Tran, L.; Liu, X.; Zhou, J.; Jin, R. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 1405–1414.
5. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [[CrossRef](#)]
6. Ma, M.; Ren, J.; Zhao, L.; Tulyakov, S.; Wu, C.; Peng, X. SMIL: Multimodal learning with severely missing modality. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 2302–2310. [[CrossRef](#)]
7. Lee, H.B.; Lee, H.; Na, D.; Kim, S.; Park, M.; Yang, E.; Hwang, S.J. Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks. In *Proceedings of the International Conference on Learning Representations*, Addis Ababa, Ethiopia, 26–30 April 2020; pp. 12850–12864.
8. Chen, M.; Zhang, F.; Zhao, Z.; Yao, J.; Zhang, Y.; Wang, Y. Probabilistic Conformal Distillation for Enhancing Missing Modality Robustness. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 10–15 December 2024; Volume 37.
9. Suo, L.; Yang, H.; Li, Q.; Yang, H.-L.; Yao, Y. A review of three-way decision: Triadic understanding, organization, and perspectives. *Int. J. Approx. Reason.* **2024**, *173*, 109268. [[CrossRef](#)]
10. Yao, Y. Three-way decision and granular computing. *Int. J. Approx. Reason.* **2018**, *103*, 107–123. [[CrossRef](#)]
11. Shi, Y.; Paige, B.; Torr, P. Variational mixture-of-experts autoencoders for multi-modal deep generative models. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
12. Liu, J.; Musialski, P.; Wonka, P.; Ye, J. Tensor completion for estimating missing values in visual data. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 208–220. [[CrossRef](#)] [[PubMed](#)]
13. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
14. Finn, C.; Abbeel, P.; Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the International Conference on Machine Learning*, Sydney, Australia, 6–11 August 2017; pp. 1126–1135.
15. Lee, H.-C.; Lin, C.-Y.; Hsu, P.-C.; Hsu, W.H. Audio feature generation for the missing-modality problem in video action recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*; IEEE: Piscataway, NJ, USA, 2019; pp. 3956–3960.
16. Balaji, Y.; Sankaranarayanan, S.; Chellappa, R. MetaReg: Towards domain generalization using meta-regularization. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Montréal, QC, Canada, 2–8 December 2018; Volume 31.
17. Li, M.; Yang, D.; Zhao, X.; Wang, S.; Wang, Y.; Yang, K.; Sun, M.; Kou, D.; Qian, Z.; Zhang, L. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 12458–12468.
18. Zolfaghari, M.; Zhu, Y.; Gehler, P.; Brox, T. CrossCLR: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 1450–1459.
19. Lian, Z.; Chen, L.; Sun, L.; Liu, B.; Tao, J. GCNet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 8419–8432. [[CrossRef](#)] [[PubMed](#)]
20. Dugas, C.; Bengio, Y.; Bélisle, F.; Nadeau, C.; Garcia, R. Incorporating second-order functional knowledge for better option pricing. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Denver, CO, USA, 27 November–2 December 2000; Volume 13.
21. Yoon, J.; Kim, T.; Dia, O.; Kim, S.; Bengio, Y.; Ahn, S. Bayesian model-agnostic meta-learning. In *Proceedings of the Annual Conference on Neural Information Processing Systems*, Montréal, QC, Canada, 2–8 December 2018; Volume 31.
22. Gordon, J.; Bronskill, J.; Bauer, M.; Nowozin, S.; Turner, R.E. Meta-learning probabilistic inference for prediction. In *Proceedings of the International Conference on Learning Representations*, New Orleans, LA, USA, 6–9 May 2019; pp. 7205–7225.
23. Liu, Z.; Shen, Y.; Lakshminarasimhan, V.B.; Liang, P.P.; Zadeh, A.B.; Morency, L.-P. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia, 15–20 July 2018; pp. 2247–2256.
24. Vielzeuf, V.; Lechervy, A.; Pateux, S.; Jurie, F. CentralNet: A multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision Workshops*, Munich, Germany, 10–13 September 2018.
25. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]

26. Kingma, D.; Ba, J. A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015; pp. 1–15.
27. Arévalo, J.; Solorio, T.; Montes-y-Gómez, M.; González, F.A. Gated multimodal networks. *Neural Comput. Appl.* **2020**, *32*, 10209–10228. [[CrossRef](#)]
28. Wu, M.; Goodman, N. Multimodal generative models for scalable weakly-supervised learning. In Proceedings of the Annual Conference on Neural Information Processing Systems, Montréal, QC, Canada, 2–8 December 2018; Volume 31.
29. Tsai, Y.-H.H.; Bai, S.; Liang, P.P.; Kolter, J.Z.; Morency, L.-P.; Salakhutdinov, R. Multimodal transformer for unaligned multimodal language sequences. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 6558–6569.
30. Hazarika, D.; Zimmermann, R.; Poria, S. MISA: Modality-invariant and specific representations for multimodal sentiment analysis. In Proceedings of the ACM International Conference on Multimedia, Virtually, 12–16 October 2020; pp. 1122–1131.
31. Yu, W.; Xu, H.; Yuan, Z.; Wu, J. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 10790–10797. [[CrossRef](#)]
32. Yuan, Z.; Li, W.; Xu, H.; Yu, W. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In Proceedings of the ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021; pp. 4400–4407.
33. Li, Z.; Xu, B.; Zhu, C.; Zhao, T. CLMLF: A contrastive learning and multi-layer fusion method for multimodal sentiment detection. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, USA, 10–15 July 2022; pp. 2282–2294.
34. Liu, R.; Zuo, H.; Lian, Z.; Schuller, B.W.; Li, H. Contrastive learning-based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Trans. Affect. Comput.* **2024**, *15*, 1856–1873. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.