Trust, but Verify: Uncertainty-Driven Evidential Multimodal Representation Learning

Yupeng Han[1], **Kai Zhang**[1*], **Xianquan Wang**[1], **Zhihong Pan**[1], **Ze Liu**[1] and **Jiyuan He**[2]
[1]State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China
[2]Meituan(China)
{yupenghan, wxqcn, panzh, liuze2120}@mail.ustc.edu.cn, kkzhang08@ustc.edu.cn, hejiyuan@meituan.com

Abstract

Effective multimodal learning in real-world scenarios depends on a nuanced treatment of uncertainty, which arises at three levels: (1) **Intrinsic Uncertainty** from modality-specific noise or ambiguity; (2) **Relational Uncertainty** due to cross-modal conflicts or redundancy; and (3) **Aggregated Uncertainty** when fusing potentially inconsistent signals. Most existing methods overlook this hierarchy, applying a single uncertainty model. We propose Adaptive Evidential Multimodal Representation Learning (**AEMRL**), a framework aligned with this multi-level view. To address intrinsic uncertainty, Disentangled Evidential Uncertainty Encoding (**D-EUE**) provides interpretable, class-aware reliability scores per modality. For relational uncertainty, Uncertainty-Conditioned Dynamic Factorization (**UDF**) uses a hypernetwork to dynamically extract complementary cues and suppress conflict. To resolve aggregated uncertainty, Adaptive Fusion & Conflict-Aware Calibration (**AFCAC**) adaptively weights evidence streams and calibrates final predictions based on detected conflicts. Extensive experiments on five diverse benchmark datasets show that AEMRL consistently enhances task accuracy, reduces calibration error, and improves robustness to noise, semantic conflict, and missing modalities.

1 Introduction

Multimodal perception lies at the core of modern AI [Baltrušaitis *et al.*, 2018; Zhang and Han, 2025; Wang *et al.*, 2026], revolutionizing how machines understand the world in high-stakes domains, from autonomous driving [Caesar *et al.*, 2020; Martin *et al.*, 2019] and medical diagnostics [Kaczmarczyk *et al.*, 2024] to human-computer interaction [Liu *et al.*, 2026; Zhang *et al.*, 2021a; Liang *et al.*, 2023]. Unlike curated lab datasets, real-world data is inherently imperfect, characterized by sensor noise, environmental ambiguity, and conflicting signals between modalities. Learning robust representations in these "in-the-wild" settings demands more than simply aligning modalities; it requires a sophisticated capacity to handle uncertainty. However, this uncertainty is not a monolithic problem. It is a multi-level phenomenon that manifests from the quality of individual data streams to the complex interactions between them. Prevailing paradigms, as illustrated in Figure 1(A), often fail because they address these challenges in a disconnected manner. A reliability-agnostic feature extractor may blindly trust a noisy input (e.g., a blurry X-ray), while a separate uncertainty analysis module correctly identifies the poor quality, but this critical insight is ignored. By failing to integrate these steps, such methods often produce brittle and dangerously overconfident models, limiting their real-world applicability. In contrast, our work proposes a systematic, hierarchical approach to resolve these uncertainties, leading to a correct and calibrated decision, as conceptualized in Figure 1(B).

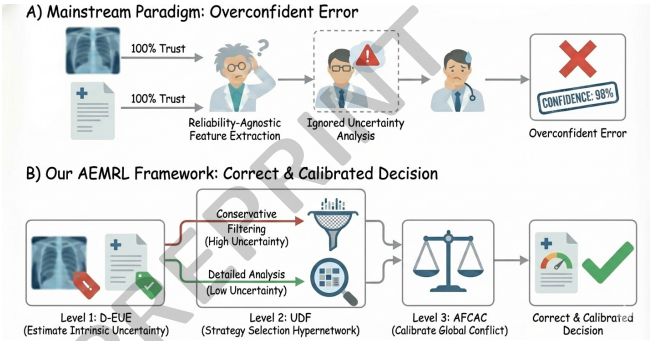

Figure 1: A conceptual illustration of AEMRL's systematic approach to multi-level uncertainty, contrasted with the disconnected process of mainstream paradigms.

The foundational level of this challenge is **Intrinsic Uncertainty**, which pertains to the reliability of each individual modality. A meme's image might be heavily compressed and artifact-ridden; a product review's text could be sarcastic or contain typos; a video's audio track may be corrupted by background noise. A dependable model must first act as a discerning critic, asking: "*Before I fuse any information, how much should I trust this single piece of evidence?*" Many current approaches [Liang *et al.*, 2023; Mai *et al.*, 2022; Yang *et al.*, 2022] sidestep this crucial first step or employ black-box reliability scores that lack interpretability, failing to identify why a modality is deemed uncertain.

*Corresponding author.

Beyond this individual assessment, a deeper **Relational Uncertainty** emerges from the interplay between modalities. This can manifest as direct conflict (e.g., an image of a peaceful landscape paired with hateful text) or as a nuanced interplay of complementarity and redundancy. For instance, a sharp product photo might merely be redundant to a glowing text review, but a blurry photo becomes critically complementary to it. This raises a more complex question: "*How can the model intelligently arbitrate between modalities, dynamically deciding whether a modality-specific cue is a critical complementary signal, a simple redundancy, or misleading noise?*" Factorization-based methods [Zhang *et al.*, 2021b; Liang *et al.*, 2023; Madaan *et al.*, 2024; Wang *et al.*, 2025] attempt to separate shared and unique features, but they do so under an implicit assumption of equal reliability, failing to adapt their factorization strategy when one modality is clearly less trustworthy than another.

Ultimately, the intrinsic and relational uncertainties identified in the preceding stages must be reconciled and combined into a single, trustworthy final decision. This brings us to the third level of our hierarchy: **Aggregated Uncertainty**. This concerns the challenge of coherently fusing multiple, potentially conflicting streams of evidence and their associated uncertainty signals. When faced with low-quality inputs or irreconcilable disagreements, a robust model should not be forced to make a confident guess. The critical question becomes: "*How can the model aggregate all known ambiguities to produce a final, well-calibrated belief that transparently reflects the total uncertainty of the system?*" Naive averaging or static fusion rules [Huang *et al.*, 2022; Zhang *et al.*, 2023] often fail at this aggregation step, producing overconfident errors in such scenarios, which is unacceptable in safety-critical applications.

To systematically address this uncertainty hierarchy, we introduce **Adaptive Evidential Multimodal Representation Learning (AEMRL)**, a framework architecturally designed to tackle each level. To address **Intrinsic Uncertainty**, we propose **Disentangled Evidential Uncertainty Encoding (D-EUE)**, a novel mechanism that provides an interpretable, class-aware reliability assessment. To manage **Relational Uncertainty**, we pioneer **Uncertainty-Conditioned Dynamic Factorization (UDF)**, which uses a hypernetwork to leverage the intrinsic uncertainty from D-EUE and dynamically generate a tailored factorization strategy. Finally, to resolve **Aggregated Uncertainty**, our **Adaptive Fusion & Conflict-Aware Calibration (AFCAC)** module intelligently integrates the purified evidence streams and scales the final belief based on inter-modal conflict, ensuring a well-calibrated and trustworthy output. Extensive experiments on five diverse benchmarks, spanning domains from medical imaging to human activity recognition, demonstrate that AEMRL consistently sets a new state-of-the-art in performance, calibration, and robustness over strong baselines.

2 Related Work

2.1 Multimodal Representation Learning

Early large-scale image–text models such as **CLIP** and **ALIGN** use dual encoders with a contrastive objective to embed paired inputs into a shared space [Radford *et al.*, 2021; Jia *et al.*, 2021]. Later "one-stream" architectures (e.g., FLAVA [Singh *et al.*, 2022], VATT [Akbari *et al.*, 2021]) introduce joint Transformers for feature fusion but assume that *shared* information is both necessary and sufficient for downstream tasks. To relax this assumption, recent work explores factorized representations. **MIB** maximizes $I(Z;Y)$ while limiting $I(Z;X)$ the learning of minimal task-relevant codes [Mai *et al.*, 2022]. **FactorCL** extends this by decomposing features into shared and modality-specific components using contrastive objectives [Liang *et al.*, 2023]. However, these approaches neither quantify epistemic uncertainty nor treat cross-modal conflict as informative. **AEMRL** advances beyond this static paradigm by introducing a synergistic, uncertainty-aware pipeline. Instead of treating uncertainty as a post-hoc fusion tool, AEMRL leverages an evidential framework as a dynamic control signal to guide feature factorization, enabling per-sample adaptive disentanglement based on modality reliability.

2.2 Uncertainty-Aware Multimodal Fusion

Quantifying predictive reliability has been extensively studied in single-modal settings, from MC-Dropout [Gal and Ghahramani, 2016] and Deep Ensembles to Dirichlet-based **Evidential Deep Learning (EDL)** [Sensoy *et al.*, 2018]. In the multimodal domain, recent methods introduce quality gates at the logit level—e.g., Quality-Aware Multimodal Fusion (QMF) [Zhang *et al.*, 2023]—or attach evidential heads per modality, as in **DDEF** (Dual-level Deep Evidential Fusion) [Shao *et al.*, 2024]. However, these approaches treat uncertainty primarily as a **fusion-stage issue**, focusing on combining modality-specific outputs while ignoring how uncertainty should shape the **feature learning process itself**. If the features are already corrupted, even optimal fusion yields subpar results. **AEMRL** fills this gap by shifting from *uncertainty-aware fusion* to *uncertainty-driven representation learning*. Its key innovation lies in the D-EUE and UDF pipeline: interpretable uncertainty signals are first generated, then used to dynamically guide feature factorization *prior* to fusion. This proactive purification ensures that inputs to the final conflict-aware calibration are already reliable, resulting in state-of-the-art performance and calibration.

3 Method

3.1 Overview

Our proposed framework, Adaptive Evidential Multimodal Representation Learning (AEMRL), is architecturally designed to systematically address the multi-level uncertainty inherent in real-world scenarios. As illustrated in Figure 2, it operates through three synergistic stages, each tailored to a specific level of our proposed uncertainty hierarchy. The following sections detail the technical implementation of each stage: first, assessing intrinsic uncertainty via **Disentangled Evidential Uncertainty Encoding (D-EUE)**; second, managing relational uncertainty via **Uncertainty-Conditioned Dynamic Factorization (UDF)**; and finally, calibrating global uncertainty via **Adaptive Fusion & Conflict-Aware Calibration (AFCAC)**.

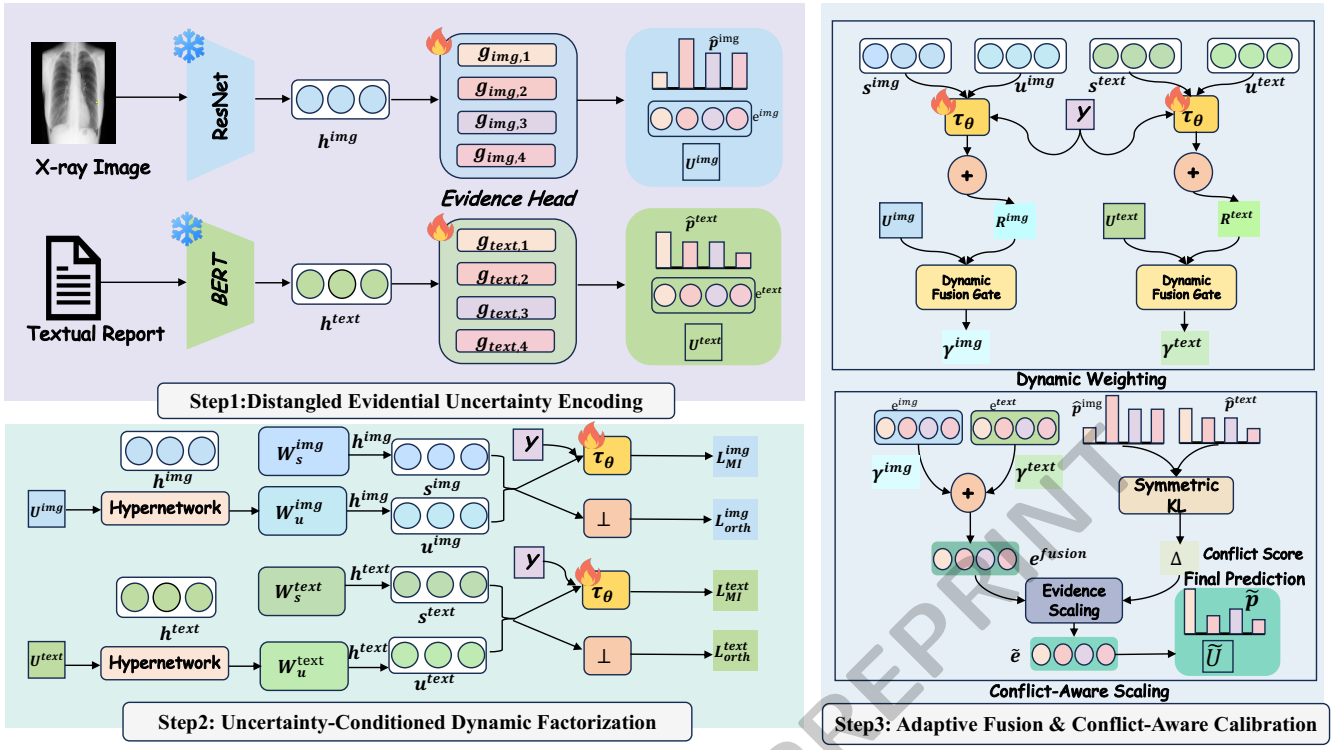

Figure 2: **Overview of the AEMRL Framework.**

3.2 Disentangled Evidential Uncertainty Encoding

To address the challenge of varying reliability and noise across different data sources in multimodal learning, we introduce the *Disentangled Evidential Uncertainty Encoding (D-EUE)* module. Traditional evidential methods [Xu *et al.*, 2024; Han *et al.*, 2022; Sensoy *et al.*, 2018] often employ a monolithic "black-box" evidence head, which obscures the connection between input features and the resulting class-level evidence, thereby limiting interpretability. Our D-EUE module enhances this paradigm by introducing a structured, parallel architecture that not only quantifies modality-specific epistemic uncertainty but also provides a more interpretable, disentangled pathway for evidence generation for each class.

D-EUE transforms each modality's feature representation into a Dirichlet distribution using a set of class-specific evidence generators, instead of a single shared network. This design enables fine-grained analysis of uncertainty sources. For each modality m, the input $\boldsymbol{x}^{(m)}$ is first encoded by a modality-specific encoder f_m to produce features $\boldsymbol{h}^{(m)}$. These are fed into a structured *evidence head* comprising C parallel generators $g_{m,c}$, one per class. Each $g_{m,c}$ (a two-layer MLP) outputs a non-negative evidence value $e_{m,c}$ dedicated to class c. The resulting evidence vector $\boldsymbol{e}^{(m)} = [e_{m,1}, \ldots, e_{m,C}]$ aggregates class-wise evidence through distinct, specialized pathways.

The evidence vector $\boldsymbol{e}^{(m)}$ is used to parameterize a Dirichlet distribution with concentration parameters $\boldsymbol{\alpha}^{(m)}$:

$$\boldsymbol{\alpha}^{(m)} = \boldsymbol{e}^{(m)} + \mathbf{1}, \tag{1}$$

where $\mathbf{1}$ is a vector of ones, ensuring $\alpha_c^{(m)} > 0$. The total evidence for modality m is $\alpha_0^{(m)} = \sum_c \alpha_c^{(m)}$, where c represents the classes. The predictive probability for modality m and the epistemic uncertainty U_m for modality m are quantified as:

$$\hat{\mathbf{p}}^{(m)} = \boldsymbol{\alpha}^{(m)} / \alpha_0^{(m)}, \quad U_m = \frac{C}{\alpha_0^{(m)}}. \tag{2}$$

A higher U_m indicates lower confidence in the information from modality m, stemming from less total evidence $\alpha_0^{(m)}$.

This explicit uncertainty U_m is then leveraged as a critical control signal in subsequent stages. Unlike prior approaches that use a single, fully coupled evidence head, D-EUE's disentangled architecture provides a more transparent and interpretable uncertainty assessment. This allows the model not just to know *that* it is uncertain but potentially *why*, by inspecting the evidence contributions from each specialized pathway. This fine-grained signal enables a more sophisticated, reliability-aware adaptation in downstream modules.

Evidence generators are trained via a loss that promotes both prediction accuracy and functional specialization. For modality m, the standard evidential loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{evd}}^{(m)} &= \mathbb{E}_{\boldsymbol{p}\sim\text{Dir}(\boldsymbol{\alpha}^{(m)})}[-\log p_y] \\ &\quad + \lambda_{\text{KL}}\,\text{KL}\Big(\text{Dir}(\boldsymbol{\alpha}^{(m)}) \,\|\, \text{Dir}(\mathbf{1})\Big).\end{aligned} \tag{3}$$

To ensure the parallel generators learn distinct functions, we add our novel disentanglement constraint. By promoting orthogonality among the generator weights ($\boldsymbol{G}_{m,c}$), this loss encourages each generator to focus on a unique subspace of features. This specialization prevents evidence meant for one

class from interfering with another, leading to a cleaner and more interpretable final uncertainty assessment. The loss is:

$$\mathcal{L}_{\text{orth}}^{(m)} = \sum_{c \neq j} \|\boldsymbol{G}_{m,c}^T \boldsymbol{G}_{m,j}\|_F^2. \tag{4}$$

The complete objective function for the D-EUE module is the weighted sum of these two components:

$$\mathcal{L}_{\text{D-EUE}}^{(m)} = \mathcal{L}_{\text{evd}}^{(m)} + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}}^{(m)}. \tag{5}$$

3.3 Uncertainty-Conditioned Dynamic Factorization

To address **relational uncertainty** under varying modality reliability, we introduce the *Uncertainty-Conditioned Dynamic Factorization (UDF)* module. While prior methods [Liang *et al.*, 2023; Yang *et al.*, 2022; Robinet *et al.*, 2024; Wang *et al.*, 2025] perform static, reliability-agnostic factorization, they apply fixed projections regardless of input quality—risking the extraction of spurious "unique" features from noise. In contrast, UDF conditions the factorization process on intrinsic uncertainty U_m, enabling dynamic adaptation of feature extraction.

At its core, UDF employs a **hypernetwork** [Ha *et al.*, 2016], HyperNet_θ, which maps a modality's uncertainty to its optimal factorization parameters, transforming factorization into a sample-aware operation. For each modality m, the two-layer MLP hypernetwork takes the scalar uncertainty U_m as input and generates the weights for the unique projection matrix $\boldsymbol{W}_u^{(m)}$. This allows the model to learn context-sensitive strategies: generating expressive projectors for reliable modalities and conservative ones for noisy inputs.

The unique feature projection matrix $\boldsymbol{W}_u^{(m)}$ is thus dynamically generated, while the shared feature projector $\boldsymbol{W}_s^{(m)}$ is a statically learned parameter to capture cross-modal commonalities. The feature $\boldsymbol{h}^{(m)}$ is decomposed into shared features $\boldsymbol{s}_m$ and unique features $\boldsymbol{u}_m$ as follows:

$$\boldsymbol{W}_u^{(m)} = \text{reshape}(\text{HyperNet}_\theta(U_m), (d_u, D)), \tag{6}$$

$$\boldsymbol{s}_m = \boldsymbol{W}_s^{(m)} \boldsymbol{h}^{(m)}, \quad \boldsymbol{u}_m = \boldsymbol{W}_u^{(m)} \boldsymbol{h}^{(m)}. \tag{7}$$

This dynamic factorization enables the model to disentangle complementary cues from conflicting noise based on real-time reliability estimates. Unlike prior methods with fixed projection matrices, UDF offers a principled, data-driven approach to managing relational uncertainty by adapting the feature space. The resulting purified representations, $\boldsymbol{s}_m$ and $\boldsymbol{u}_m$, form a cleaner foundation for the final fusion stage.

The UDF module is trained by minimizing a hybrid loss that combines an information-theoretic objective with a decorrelation term. It enhances the relevance of shared features while suppressing irrelevant or unique ones through the following criterion:

$$\mathcal{L}_{\text{MI}}^{(m)} = -I_{\text{LB}}(\boldsymbol{s}_m; Y) + \beta\, I_{\text{UB}}(\boldsymbol{u}_m; Y \mid \boldsymbol{s}_m). \tag{8}$$

To estimate the mutual information lower bound, I_{LB}, we employ the widely used Donsker-Varadhan (DV) estimator [Belghazi *et al.*, 2018]. This estimator utilizes a critic network, $T_\theta(\cdot, \cdot)$, to learn a lower bound on the KL-divergence between the joint and the product of marginal distributions. The DV lower bound is formulated as:

$$I_{\text{LB}}(\boldsymbol{z}; Y) = \mathbb{E}_{p(\boldsymbol{z},y)}[T_\theta(\boldsymbol{z}, y)] - \log \mathbb{E}_{p(\boldsymbol{z})p(y)}[e^{T_\theta(\boldsymbol{z},y)}]. \tag{9}$$

We apply this estimator to both the shared features ($I_{\text{LB}}(\boldsymbol{s}_m; Y)$) and, in conditional form, to the unique features. For the upper bound I_{UB}, we use the corresponding dual formulation. The critic T_θ is a two-layer MLP. This information-theoretic objective encourages the model to encode all task-relevant information into the shared representation $\boldsymbol{s}_m$.

To further push the shared and unique channels into complementary sub-spaces, we add a soft decorrelation term:

$$\mathcal{L}_{\text{orth}}^{(m)} = \left\|\boldsymbol{s}_m^\top \boldsymbol{u}_m\right\|_2^2. \tag{10}$$

3.4 Adaptive Fusion & Conflict-Aware Calibration

Leveraging the purified features from UDF and the initial evidence from D-EUE, the final stage addresses the challenge of producing a well-calibrated **global decision** via the *Adaptive Fusion & Conflict-Aware Calibration (AFCAC)* module. Traditional methods often use naive averaging or static fusion rules [Zhang *et al.*, 2023; Peng *et al.*, 2022], which can lead to overconfident errors when faced with conflicting or low-quality evidence. AFCAC, instead, intelligently integrates the available information and calibrates the final prediction's confidence based on inter-modal agreement.

The core of AFCAC is a two-step process of dynamic weighting and conflict-aware scaling. First, it computes a fusion weight γ_m for each modality that reflects both its task relevance and its reliability. The relevance score, R_m, quantifies the total task-relevant information from modality m. We define it as the sum of the mutual information lower bounds for its shared and unique features, as estimated by the critic networks T_θ in the UDF stage:

$$R_m = I_{\text{LB}}(\boldsymbol{s}_m; Y) + I_{\text{LB}}(\boldsymbol{u}_m; Y \mid \boldsymbol{s}_m). \tag{11}$$

The reliability is given by $(1 - U_m)$ from the D-EUE stage. These two components are combined and normalized across all M modalities to produce the final dynamic weight γ_m:

$$\gamma_m = \frac{R_m \cdot (1 - U_m)}{\sum_{k=1}^{M} R_k \cdot (1 - U_k)}. \tag{12}$$

The modality-specific evidences $\boldsymbol{e}^{(m)}$ are then fused via this adaptive weighted sum:$\boldsymbol{e}_{\text{fusion}} = \sum_{m=1}^{M} \gamma_m \boldsymbol{e}^{(m)}$.

Second, the module explicitly detects inter-modal conflict to prevent overconfident predictions. We measure the disagreement by calculating the symmetric Kullback-Leibler (KL) divergence between the predicted probability distributions, $\hat{\mathbf{p}}^{(v)}$ and $\hat{\mathbf{p}}^{(w)}$, of each pair of modalities. The total conflict score, Δ, is the sum of all pairwise disagreements:

$$\Delta = \sum_{v<w} \left[\text{KL}\left(\hat{\mathbf{p}}^{(v)} \,\|\, \hat{\mathbf{p}}^{(w)}\right) + \text{KL}\left(\hat{\mathbf{p}}^{(w)} \,\|\, \hat{\mathbf{p}}^{(v)}\right)\right]. \tag{13}$$

This conflict score is then used to scale the fused evidence:

$$\boldsymbol{e}_{\text{scaled}} = \boldsymbol{e}_{\text{fusion}} / (1 + \kappa\Delta), \tag{14}$$

where κ is a tunable hyperparameter that controls the sensitivity to conflict. The final Dirichlet parameters are then computed as $\boldsymbol{\alpha}_{\text{final}} = \boldsymbol{e}_{\text{scaled}} + \mathbf{1}$.

This explicit calibration ensures the model's final output reflects the cumulative uncertainty from the entire pipeline. Unlike prior approaches that might overlook conflicts or use static fusion rules, AFCAC provides a robust mechanism to manage disagreements and prevent overconfidence. It allows the model to "fail gracefully" by reporting high **aggregated uncertainty** when the collective evidence is contradictory.

The entire AEMRL framework is trained with a composite loss that combines module-specific objectives, anchored by the final classification loss. The overall objective is:

$$\mathcal{L}_{\text{AEMRL}} = \mathcal{L}_{\text{CE}}(\boldsymbol{p}_{\text{final}}, y) + \sum_{m=1}^{M} \left(\lambda_{\text{evd}} \mathcal{L}_{\text{D-EUE}}^{(m)} + \lambda_{\text{udf}} (\mathcal{L}_{\text{MI}}^{(m)} + \mathcal{L}_{\text{orth}}^{(m)}) \right). \tag{15}$$

Here, $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss between the final prediction $\boldsymbol{p}_{\text{final}}$ and the one-hot label y. The hyperparameters λ_{evd} and λ_{udf} weight the uncertainty and factorization objectives. This unified loss guides all components to jointly manage the full uncertainty hierarchy.

4 Experiment

4.1 Experimental Setup

Datasets. To demonstrate AEMRL's ability to generalize across diverse web scenarios—including safety-critical settings, sensor fusion under challenging conditions, and multimedia understanding—we evaluate on five complementary benchmarks: MIMIC-CXR [Johnson *et al.*, 2019], Drive&Act [Martin *et al.*, 2019], VGG-Sound [Chen *et al.*, 2020], Hateful Memes [Kiela *et al.*, 2020] and MM-IMDb [Arevalo *et al.*, 2017]. Full dataset statistics and preprocessing details are in Supplementary Material.

Baseline. We compare AEMRL against a broad range of SOTA methods that reflect key directions in multimodal representation learning: (1) **Conventional alignment and fusion**: CLIP [Radford *et al.*, 2021], ALIGN [Jia *et al.*, 2021], FLAVA [Singh *et al.*, 2022], and VATT [Akbari *et al.*, 2021]; (2) **Shared–unique factorization**: MMIB [Mai *et al.*, 2022], FactorCL [Liang *et al.*, 2023], I2M2 [Madaan *et al.*, 2024], and DISENTANGLEDSSL [Wang *et al.*, 2025]; (3) **Uncertainty-aware fusion**: DDEF [Shao *et al.*, 2024], Contextual Discounting (CD) [Huang *et al.*, 2022], QMF [Zhang *et al.*, 2023], OGM-GE [Peng *et al.*, 2022], and EAU [Gao *et al.*, 2024]; (4) **Missing-modality adaptation**: ModDrop [Neverova *et al.*, 2015] and MLA [Zhang *et al.*, 2024]. More details are provided in the Supplementary Material.

Implementation Details. All models are implemented in PyTorch and trained on 4 A100 GPUs. For each dataset, we use standard pretrained backbones—ResNet-50 [He *et al.*, 2016] for images, 3D ResNet-50 [Hara *et al.*, 2018] for video, BERT-base [Devlin *et al.*, 2019] for text, and CNN14 [Wang *et al.*, 2022] for audio/IMU—and fine-tune end-to-end. We adopt canonical task-specific metrics: AUROC (MIMIC-CXR), F1 (Drive&Act), Top-1 (VGGSound), Accuracy (Hateful Memes), and mAP (MM-IMDb), and uniformly report Expected Calibration Error (ECE). Optimization uses AdamW [Kingma and Ba, 2014] with a cosine learning-rate schedule and batch size of 64 per GPU. Results are averaged over five seeds (0–4). Key module configurations are as follows. The D-EUE module uses lightweight, class-specific generators, each a two-layer MLP with hidden size 128. The UDF hypernetwork is also a two-layer MLP with hidden size 64 and GELU activation. The trade-off parameter β in the mutual information objective (Eq. 8) is set to 0.5. Loss balancing hyperparameters, selected via grid search on a validation set, are $\lambda_{\text{evd}} = 1.0$, $\lambda_{\text{udf}} = 0.5$, and $\lambda_{\text{orth}} = 0.1$ across all datasets. The conflict scaling factor in AFCAC is $\kappa = 2.5$. Full experimental details, including dataset-specific preprocessing, splits, and hyperparameter listings, are provided in the Supplementary Material.

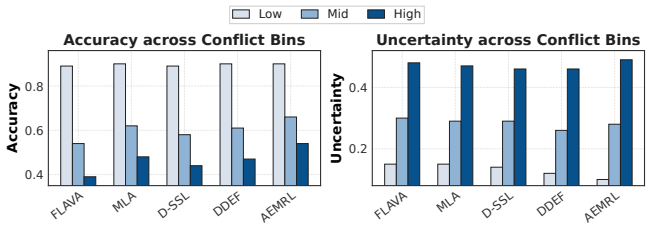

Figure 3: Robustness to Semantic Conflict.

4.2 Main Results

As Table 1 illustrates, our framework consistently establishes a new state-of-the-art on the primary performance metrics across all tasks, but the nature and magnitude of these gains vary in ways that directly support our central thesis.

Firstly, AEMRL's largest gains appear on datasets with unreliable modalities, notably Drive&Act and VGGSound. On Drive&Act, which contains noisy IMU sensor data, AEMRL achieves an F1 of 0.732, exceeding DISENTANGLEDSSL by 2.9 points. On VGGSound, where visual cues are weakly aligned with audio, the Top-1 accuracy improves by 2.3 points (0.380 vs. 0.357). These results strongly validate our Uncertainty-Conditioned Dynamic Factorization (UDF): unlike static methods that uniformly handle modality-specific features, AEMRL adaptively suppresses noisy IMU signals or irrelevant video frames, preventing them from degrading the final representation. Secondly, On cleaner datasets such as MIMIC-CXR, AEMRL shows smaller accuracy gains but a clear advantage in calibration. This indicates that even with relatively clean modalities, the combined reasoning of D-EUE and AFCAC yields a more trustworthy model—one that not only improves accuracy but also recognizes subtle ambiguities, an essential property for high-stakes domains such as medical diagnosis. Finally, On tasks involving semantic ambiguity, such as Hateful Memes and MM-IMDb, AEMRL consistently outperforms all baselines. The accuracy gain on Hateful Memes highlights its ability to handle not only signal-level noise but also high-level semantic con-

Method	MIMIC-CXR		Drive&Act		VGGSound		Hateful Memes		MM-IMDb	
	AUROC↑	ECE↓	F1↑	ECE↓	Top-1↑	ECE↓	Acc↑	ECE↓	mAP↑	ECE↓
CLIP	0.812	0.186	0.664	0.213	0.322	0.157	0.617	0.174	0.439	0.176
FLAVA	0.825	0.172	0.678	0.198	0.338	0.143	0.629	0.160	0.454	0.165
ALIGN	0.820	0.174	0.670	0.205	0.331	0.147	0.623	0.168	0.443	0.168
VATT	0.829	0.169	0.681	0.192	0.342	0.141	0.632	0.158	0.458	0.162
MMIB	0.831	0.169	0.685	0.191	0.344	0.139	0.636	0.152	0.459	0.152
DISENTANGLEDSSL	0.849	0.152	0.703	0.169	0.355	0.130	0.642	0.152	0.470	0.113
FactorCL	0.838	0.161	0.693	0.185	0.351	0.134	0.640	0.150	0.466	0.144
I2M2	0.841	0.156	0.694	0.187	0.348	0.139	0.638	0.155	0.471	0.138
DDEF	0.842	0.158	0.697	0.179	0.357	0.129	0.637	0.151	0.471	0.113
CD	0.839	0.160	0.693	0.182	0.353	0.131	0.635	0.153	0.468	0.140
QMF	0.834	0.167	0.689	0.186	0.346	0.135	0.631	0.157	0.457	0.146
OGM-GE	0.828	0.171	0.698	0.183	0.349	0.132	0.638	0.153	0.463	0.141
EAU	0.850	0.161	0.721	0.171	0.352	0.130	0.641	0.142	0.474	0.115
ModDrop	0.820	0.180	0.677	0.200	0.335	0.145	0.625	0.162	0.450	0.168
MLA	0.845	0.160	0.695	0.182	0.354	0.131	0.639	0.152	0.469	0.139
AEMRL (Ours)	**0.865**	**0.109**	**0.732**	**0.117**	**0.380**	**0.085**	**0.659**	**0.106**	**0.492**	**0.096**

Table 1: Complete-modality performance on five multimodal benchmarks. The best result per column is in bold, and the second-best is underlined. (↑: higher is better, ↓: lower is better).

flict. Through UDF's adaptive factorization and AFCAC's explicit conflict scaling, the model learns to correctly resolve cases where one modality alters the interpretation of another.

4.3 Ablation Studies

Table 2 details our ablation study across five benchmarks, validating the necessity and synergy of the three-level uncertainty hierarchy in AEMRL. First, removing all evidential components to create a deterministic baseline severely degrades performance, elevating Expected Calibration Error, ECE, by 40 to 50% on Drive&Act and MIMIC-CXR. Substituting the decoupled D-EUE with a standard monolithic EUE head similarly impairs both accuracy and calibration, confirming the advantage of disentangled uncertainty signals. Next, replacing the uncertainty-directed factorization, UDF, with static factorization triggers the largest overall performance drop, proving that processing must dynamically adapt to modality reliability. When comparing the UDF hypernetwork against scalar and vector gating on Drive&Act, the full hypernetwork strictly outperforms both alternatives. This demonstrates its superior capacity to capture fine-grained reliability variations, particularly under noise. Finally, ablating the conflict scaling mechanism within AFCAC reduces AUROC by a mere 0.006 on MIMIC-CXR but substantially worsens calibration, increasing ECE by 15 to 25% across all benchmarks. Additionally, substituting dynamic weighting with naive averaging degrades overall performance, proving that reliability-aware fusion remains indispensable even following UDF purification.

4.4 Robustness to Modality Conflict

To evaluate AEMRL's ability to handle high-level semantic disagreement—a central challenge in Web content analysis—we conducted a conflict stress test on the Hateful Memes benchmark. The test set was partitioned into three bins of increasing conflict, based on the pre-fusion symmetric KL-divergence between image and text modalities. We then measured accuracy and reported uncertainty for AEMRL and key baselines within each bin. As shown in Figure 3, AEMRL demonstrates superior resilience to inter-modal conflict. In the left panel, all models show degraded accuracy as conflict increases, but AEMRL degrades more gracefully. In the high-conflict bin, it maintains 54.1% accuracy—an 11.8-point gain over FactorCL, which drops by over 46 points—indicating AEMRL's stronger capacity to arbitrate between contradictory signals. The right panel reveals the source of this robustness. Although all models report higher uncertainty under conflict, AEMRL's increase is more pronounced and calibrated. In the high-conflict regime, its average uncertainty rises to 0.495, accurately reflecting input ambiguity. This behavior results from the AFCAC, which detects high KL-divergence and down-weights the fused evidence, preventing overconfident errors. These results confirm AEMRL's principled and robust handling of conflicting inputs.

4.5 Robustness to Signal-Level Degradation

The results in Figure 4 validate AEMRL's adaptive mechanisms. As shown in Panel (a), the static model's performance collapses under image degradation, with AUROC dropping by over 15 points. In contrast, AEMRL degrades gracefully and maintains high AUROC even under severe blur ($\sigma = 4$), effectively ignoring the corrupted image and relying on the clean textual report. Panels (b)–(d) reveal the internal mechanisms behind this robustness. Panel (b) shows that D-EUE acts as a reliable quality sensor: as blur increases, the intrinsic uncertainty U_{img} rises predictably. Panel (c) provides di-

Method	MIMIC-CXR		Drive&Act		VGGSound		Hateful Memes		MM-IMDb	
	AUROC↑	ECE↓	F1↑	ECE↓	Top-1↑	ECE↓	Acc↑	ECE↓	mAP↑	ECE↓
AEMRL (Full Model)	**0.865**	**0.109**	**0.732**	**0.117**	**0.380**	**0.085**	**0.659**	**0.106**	**0.492**	**0.096**
① **w/o Uncertainty Assessment**										
w/o D-EUE (use Standard EUE)	0.854	0.125	0.718	0.121	0.371	0.095	0.649	0.115	0.481	0.105
w/o Uncertainty (deterministic)	0.845	0.152	0.704	0.178	0.360	0.128	0.635	0.141	0.470	0.131
② **w/o Adaptive Factorization**										
w/o UDF (use Static Factorization)	0.849	0.131	0.701	0.145	0.359	0.134	0.638	0.128	0.472	0.120
w/o HyperNet (use Vector Gating)	<u>0.856</u>	<u>0.119</u>	0.714	0.125	0.370	0.099	0.647	0.113	0.480	0.107
w/o HyperNet (use Scalar Gating)	0.851	0.126	0.708	0.133	0.365	0.115	0.641	0.121	0.475	0.115
③ **w/o Final Calibration**										
w/o Conflict Scaling	<u>0.859</u>	0.128	<u>0.724</u>	0.129	<u>0.372</u>	0.105	<u>0.652</u>	0.123	<u>0.486</u>	0.114
w/o Dynamic Weights (use mean)	0.852	0.124	0.711	0.138	0.368	0.101	0.645	0.119	0.478	0.109

Table 2: Comprehensive ablation study across all five benchmarks. The best result per column is in bold, and the second-best is underlined.

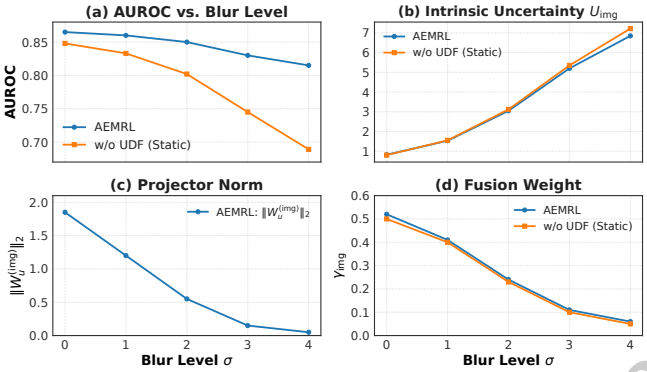

Figure 4: AEMRL's Robustness and Internal Dynamics under Controlled Degradation. We progressively apply Gaussian blur (parameterized by σ) to the image modality in the MIMIC-CXR test set and track four key metrics.

rect evidence of UDF's adaptation. As blur increases, the spectral norm of the unique feature projector $\|W_u^{\text{img}}\|_2$ decreases, indicating that the hypernetwork learns to suppress noisy modality-specific features. Panel (d) shows that this assessment propagates to fusion: the image modality's weight γ_{img} is progressively down-weighted. This coordinated response from uncertainty assessment to adaptive factorization to fusion is the key to AEMRL's robustness. The experiment confirms that AEMRL is not only performant on average, but also equipped with a principled, verifiable mechanism for handling signal degradation.

4.6 Robustness to Missing Modalities

We evaluate zero-shot robustness to missing modalities on Drive&Act. All models are trained only on complete data, while at inference the RGB stream is randomly dropped with probability η; F1 and ECE are then measured. As shown in Figure 5(a), all methods degrade as η increases, but AEMRL remains the most stable and achieves the highest F1. At $\eta = 80\%$, AEMRL drops by only 6.1 points, while MLA drops by 5.7 points from a much lower baseline and DISENTANGLEDSSL collapses by 14 points. This robustness

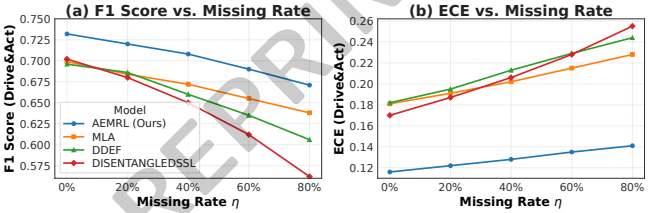

Figure 5: Robustness to Missing Modalities on Drive&Act.

comes from AEMRL's uncertainty-aware design, where missing modalities are treated as extreme intrinsic uncertainty and their influence is suppressed by D-EUE, UDF, and AFCAC. Thus, unreliable or absent streams contribute little to the final decision. This design avoids overfitting to complete-modality assumptions. It also enables the model to adapt to incomplete inputs without retraining. Figure 5(b) further shows that AEMRL maintains low and stable ECE, whereas baselines become poorly calibrated, with DISENTANGLEDSSL increasing by over 50%. These results confirm that AEMRL preserves both accuracy and calibration under incomplete multimodal inputs.

5 Conclusion

We introduced AEMRL, a novel framework that systematically addresses the multi-level hierarchy of uncertainty inherent in real-world multimodal data. AEMRL's core innovation is an uncertainty-driven pipeline where reliability assessment (D-EUE) dynamically controls feature factorization (UDF), followed by a conflict-aware calibration (AFCAC). This synergistic design moves beyond the limitations of prior work, which treats uncertainty and feature learning as disconnected processes. Comprehensive experiments validated our approach. AEMRL established a new state-of-the-art in both performance and calibration across five diverse benchmarks. Through a series of targeted ablations and stress tests, we provided direct empirical evidence that our framework's superior robustness stems from its principled, multi-level handling of semantic conflict, signal degradation, and missing modalities.

Acknowledgements

This research was partially supported by the National Natural Science Foundation of China (Grants No.62406303), Anhui Province Science and Technology Innovation Project (202423k09020010), Preliminary Research Pilot Project of the Chinese Academy of Sciences (XDA0490203) and Meituan.

References

[Akbari *et al.*, 2021] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in neural information processing systems*, 34:24206–24221, 2021.

[Arevalo *et al.*, 2017] John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

[Baltrušaitis *et al.*, 2018] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[Belghazi *et al.*, 2018] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR, 2018.

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[Chen *et al.*, 2020] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Gal and Ghahramani, 2016] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[Gao *et al.*, 2024] Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26876–26885, 2024.

[Ha *et al.*, 2016] David Ha, Andrew Dai, and Quoc V. Le. Hypernetworks, 2016.

[Han *et al.*, 2022] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):2551–2566, 2022.

[Hara *et al.*, 2018] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Huang *et al.*, 2022] Ling Huang, Thierry Denoeux, Pierre Vera, and Su Ruan. Evidence fusion with contextual discounting for multi-modality medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 401–411. Springer, 2022.

[Jia *et al.*, 2021] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[Johnson *et al.*, 2019] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[Kaczmarczyk *et al.*, 2024] Robert Kaczmarczyk, Theresa Isabelle Wilhelm, Ron Martin, and Jonas Roos. Evaluating multimodal ai in medical diagnostics. *npj Digital Medicine*, 7(1):205, 2024.

[Kiela *et al.*, 2020] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Liang *et al.*, 2023] Paul Pu Liang, Zihao Deng, Martin Q Ma, James Y Zou, Louis-Philippe Morency, and Ruslan Salakhutdinov. Factorized contrastive learning: Going beyond multi-view redundancy. *Advances in Neural Information Processing Systems*, 36:32971–32998, 2023.

[Liu *et al.*, 2026] Shuochen Liu, Pengfei Luo, Chao Zhang, Yuhao Chen, Haotian Zhang, Qi Liu, Xin Kou, Tong Xu, and Enhong Chen. Look as you think: Unifying reasoning and visual evidence attribution for verifiable document rag via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 32159–32167, 2026.

[Madaan *et al.*, 2024] Divyam Madaan, Taro Makino, Sumit Chopra, and Kyunghyun Cho. Jointly modeling inter- & intra-modality dependencies for multi-modal learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Mai *et al.*, 2022] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2022.

[Martin *et al.*, 2019] Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, and Rainer Stiefelhagen. Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2801–2810, 2019.

[Neverova *et al.*, 2015] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2015.

[Peng *et al.*, 2022] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8238–8247, 2022.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Robinet *et al.*, 2024] Lucas Robinet, Ahmad Berjaoui, Ziad Kheil, and Elizabeth Cohen-Jonathan Moyal. Drim: Learning disentangled representations from incomplete multimodal healthcare data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 163–173. Springer, 2024.

[Sensoy *et al.*, 2018] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018.

[Shao *et al.*, 2024] Zhimin Shao, Weibei Dou, and Yu Pan. Dual-level deep evidential fusion: Integrating multimodal information for enhanced reliable decision-making in deep learning. *Information Fusion*, 103:102113, 2024.

[Singh *et al.*, 2022] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650, 2022.

[Wang *et al.*, 2022] Zhepei Wang, Cem Subakan, Xilin Jiang, Junkai Wu, Efthymios Tzinis, Mirco Ravanelli, and Paris Smaragdis. Learning representations for new sound classes with continual self-supervised learning. *IEEE Signal Processing Letters*, 29:2607–2611, 2022.

[Wang *et al.*, 2025] Chenyu Wang, Sharut Gupta, Xinyi Zhang, Sana Tonekaboni, Stefanie Jegelka, Tommi Jaakkola, and Caroline Uhler. An information criterion for controlled disentanglement of multimodal data. In *The Thirteenth International Conference on Learning Representations*, 2025.

[Wang *et al.*, 2026] Xianquan Wang, Zhaocheng Du, Huibo Xu, Shukang Yin, Yupeng Han, Jieming Zhu, Kai Zhang, and Qi Liu. Personalized visual content generation in conversational systems. *Advances in Neural Information Processing Systems*, 38:34571–34602, 2026.

[Xu *et al.*, 2024] Cai Xu, Yilin Zhang, Ziyu Guan, and Wei Zhao. Trusted multi-view learning with label noise. *arXiv preprint arXiv:2404.11944*, 2024.

[Yang *et al.*, 2022] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pages 1642–1651, 2022.

[Zhang and Han, 2025] Kai Zhang and Yupeng Han. Harnessing commonsense: Llm-driven knowledge integration for fine-grained sentiment analysis. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4106–4116, 2025.

[Zhang *et al.*, 2021a] Kai Zhang, Qi Liu, Hao Qian, Biao Xiang, Qing Cui, Jun Zhou, and Enhong Chen. Eatn: An efficient adaptive transfer network for aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):377–389, 2021.

[Zhang *et al.*, 2021b] Kai Zhang, Hao Qian, Qing Cui, Qi Liu, Longfei Li, Jun Zhou, Jianhui Ma, and Enhong Chen. Multi-interactive attention network for fine-grained feature learning in ctr prediction. In *Proceedings of the 14th ACM international conference on web search and data mining*, pages 984–992, 2021.

[Zhang *et al.*, 2023] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023.

[Zhang *et al.*, 2024] Xiaohui Zhang, Jaehong Yoon, Mohit Bansal, and Huaxiu Yao. Multimodal representation learning by alternating unimodal adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27456–27466, 2024.