WHEN MODALITIES DISAGREE: LEARNING ADAPTIVE TASK-AWARE MULTIMODAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Human-centered multimodal understanding integrates text, audio, and visual cues to infer human sentiment and intent. However, semantic discrepancies, conflicting evidence, and cross-modal redundancy make it challenging to identify task-relevant information and determine appropriate modality contributions during fusion. Predictive disagreement alone cannot determine which evidence should guide the final decision. We introduce **InfoMind**, which formulates multimodal fusion as inference of a sample-specific latent task state. Using text-conditioned shared-private decomposition and asymmetric information-bottleneck regularization, InfoMind constructs compact, comparable task-level evidence and derives pairwise predictive disagreement. A shared-context-initialized latent state is refined once using these signals to determine modality contributions, which guide fusion of the full-width shared and private representations retained from the decomposition stage. Experiments on CMU-MOSI, CMU-MOSEI, MIntRec, and MIntRec2.0 demonstrate its effectiveness and cross-task applicability across sentiment prediction and intent recognition.

1 INTRODUCTION

Humans rarely infer a person's affective state or intent from a single source of evidence. Instead, they integrate language, vocal prosody, facial expressions, and other perceptual cues when making such inferences. Research on multisensory perception suggests that cue contributions are not fixed and that observers rely more on sources that are more informative and less uncertain (Ernst & Banks, 2002). Moreover, when sensory observations are inconsistent, perceptual systems do not necessarily integrate them indiscriminately, but can reason about whether the observations should be combined under a common interpretation (Körding et al., 2007; Shams & Beierholm, 2010). Similar challenges arise in human communication. Positive wording can be accompanied by flat or negative prosody, while facial expressions may further modify the implied meaning. Prosody is particularly important for conveying non-literal intentions such as sarcasm (Cheang & Pell, 2008). These observations raise a fundamental computational question about *how a learning system should organize and integrate heterogeneous evidence when different modalities support conflicting interpretations.*

Multimodal representation learning must handle shared, complementary, redundant, and potentially conflicting information across modalities. Recent emotion-conflict and label-divergence studies show that such discrepancies can substantially affect multimodal prediction (Han et al., 2025; Du et al., 2025). Importantly, feature discrepancy does not necessarily imply task-level disagreement, and predictive disagreement alone does not determine how modalities should be weighted during fusion. Inspired by the cognitive view that heterogeneous cues are interpreted in relation to an underlying state, we formulate multimodal fusion as task-state inference. A sample-specific latent state jointly contextualizes task-level evidence and predictive disagreement before the model determines modality contributions. Figure 1 illustrates this process.

Existing work studies information-theoretic representation learning (Xiao et al., 2024; Wu et al., 2025), sample-level interaction modeling (Yang et al., 2025; 2026), and adaptive or conflict-aware fusion (Cao et al., 2024; Wang et al., 2026; Zou & Wei, 2026). Motivated by these advances, we ask *whether predictive disagreement should directly determine fusion weights or instead serve as context for updating an evidence-dependent task state.*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

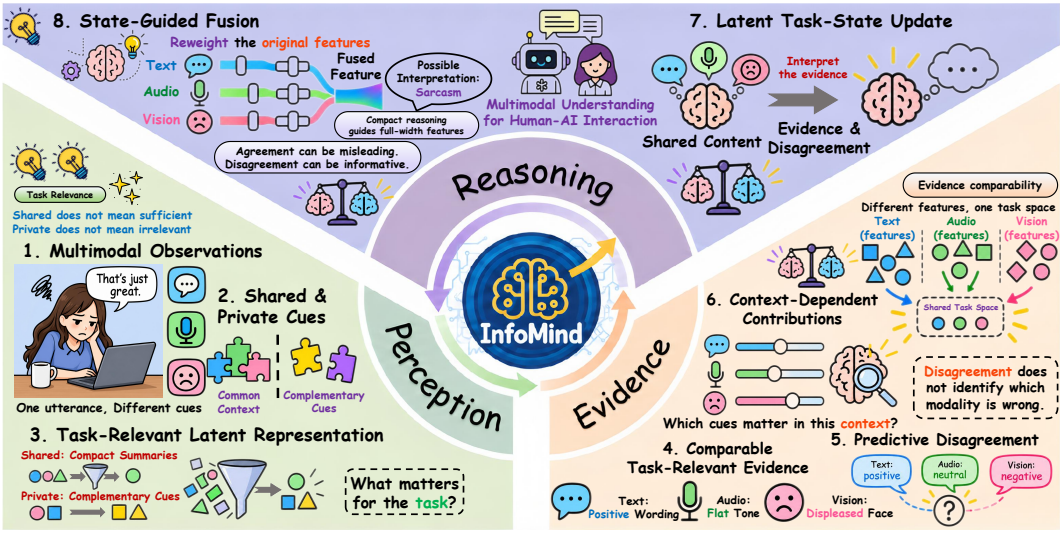


Figure 1: Conceptual motivation for adaptive multimodal fusion. Different modalities may provide shared, complementary, or conflicting cues for the same task. Their relevance depends on both the downstream task and the context of each sample, motivating a latent task state that organizes multimodal evidence before fusion.

To investigate this question, we introduce **InfoMind**, which formulates multimodal fusion as sample-specific latent task-state inference. InfoMind uses text-conditioned decomposition and asymmetric information-bottleneck regularization to construct compact, modality-associated evidence in a common task space, from which pairwise predictive disagreement is computed. Shared context initializes the latent state, which is updated once using the evidence and disagreement to determine normalized modality contributions. These contributions are converted into gates that modulate the full-width shared and private representations retained from the decomposition stage for final fusion and prediction. Our main contributions are as follows:

- We formulate adaptive multimodal fusion as a task-aware evidence reasoning problem. Inspired by cognitive principles of context-dependent cue integration, we distinguish cross-modal feature discrepancy from predictive disagreement and introduce an intermediate latent task state that contextualizes potentially conflicting evidence.
- We develop InfoMind, which combines shared-private semantic decomposition, asymmetric information-bottleneck regularization, task-level evidence construction, and a single disagreement-aware state-refinement step to guide sample-specific multimodal fusion.
- We evaluate InfoMind across continuous sentiment analysis and fine-grained multimodal intent recognition, together with controlled ablations and representation analyses that examine the contributions of its major components.

2 Related Work

Multimodal representation learning. Information-bottleneck methods compress multimodal representations to preserve task-relevant information (Kawaguchi et al., 2023; Xiao et al., 2024; Wu et al., 2025), while other work characterizes redundancy, uniqueness, and synergy across modalities (Yang et al., 2025; 2026). Shared–private methods separate cross-modal commonality from modality-specific information (Lv et al., 2025). InfoMind constructs compact task-space evidence and predictive disagreement from the decomposed streams, uses them to refine a shared-context-initialized latent state, and applies the resulting modality contributions to the full-width shared and private representations.

Adaptive Multimodal Fusion. Modality competition and unimodal bias reveal unequal contributions under joint multimodal learning (Huang et al., 2022; Zhang et al., 2024b). Recent methods adapt fusion according to modality quality, confidence, or sample-specific reliability (Cao et al.,

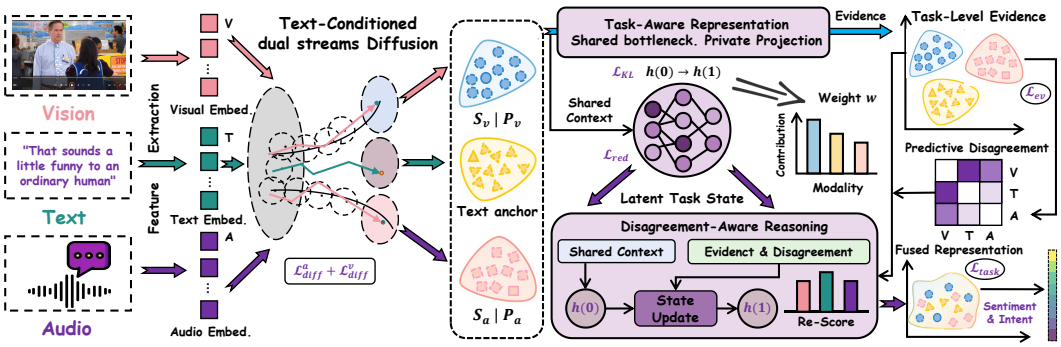


Figure 2: Overview of InfoMind. Text-conditioned diffusion decomposes acoustic and visual representations into shared and residual streams. Compact task-aware representations yield modality-associated evidence and auxiliary predictions used to compute pairwise disagreement. Shared context initializes a latent task state, which is updated once with these signals to determine modality contributions. These contributions gate the full-width shared and private representations retained from decomposition for final prediction.

2024; Fang et al., 2026; Jiang et al., 2026; He et al., 2026). These approaches motivate dynamic modality weighting, whereas InfoMind interprets task-level evidence and predictive disagreement through an intermediate latent task state before determining modality contributions.

Conflict-Aware Multimodal Learning. Recent methods model multimodal inconsistency through conflict-dependent reconstruction or weighting (Wang et al., 2026; Zou & Wei, 2026; Du et al., 2025). DRKF (Jiang et al., 2025) preserves discriminative emotional inconsistency, while WETR (Liu et al., 2026) treats emotion as a latent state for training-free token reweighting. InfoMind instead derives modality contributions from a sample-specific latent task state initialized from shared context and refined with comparable task-space evidence and predictive disagreement.

3 METHODS

3.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

Problem Formulation. Let $\mathcal{M} = \{t, a, v\}$ denote the set of textual, acoustic, and visual modalities, with $M = |\mathcal{M}| = 3$ denoting the number of modalities. After temporal alignment and modality-specific encoding and projection, each sample is represented by $\mathbf{T}, \mathbf{A}, \mathbf{V} \in \mathbb{R}^{L\times d}$, where L and d denote the aligned sequence length and common feature dimension, respectively. The target satisfies $Y \in \mathbb{R}$ for regression and $Y \in \{1, \dots, C\}$ for C-class classification, with $\widehat{Y} = F_\theta(\mathbf{T}, \mathbf{A}, \mathbf{V})$. Because the modalities may provide shared, complementary, or conflicting evidence about Y, we seek compact task-relevant representations that retain useful modality-specific cues and support evidence comparison in a common task space.

Framework Overview. As illustrated in Figure 2, InfoMind first decomposes each non-textual modality into a text-predictable shared stream and a residual stream. A compact reasoning branch applies asymmetric bottleneck compression and task-conditioned dependence regularization to construct modality-associated evidence and pairwise predictive disagreement. An aggregated shared-context representation initializes a sample-specific latent state, which is updated once using the evidence and disagreement to determine normalized modality contributions. These contributions are converted into gates that modulate the full-width shared and private representations retained from the decomposition stage for final fusion and prediction.

3.2 DUAL-STREAM DIFFUSION FOR SHARED-PRIVATE SEMANTIC DECOMPOSITION

Conditional Shared-Private Decomposition. Writing $\mathbf{H}_a = \mathbf{A}$ and $\mathbf{H}_v = \mathbf{V}$, we define the ideal text-predictable component and residual for each $m \in \{a, v\}$ as $\mathbf{S}^\star_m(\mathbf{T}) = \mathbb{E}[\mathbf{H}_m|\mathbf{T}]$ and the corresponding residual as $\mathbf{P}^\star_m = \mathbf{H}_m - \mathbf{S}^\star_m(\mathbf{T})$. Under the squared-error criterion, the conditional expectation is the optimal predictor of the non-textual representation from textual context:

$$\mathbf{S}^\star_m = \arg\min_f \mathbb{E}\left[\|\mathbf{H}_m - f(\mathbf{T})\|_2^2\right], \tag{1}$$

where $m \in \{a, v\}$. Consequently, $\mathbb{E}[\mathbf{P}_m^\star|\mathbf{T}] = \mathbf{0}$. Under this idealized formulation, the shared stream corresponds to the conditional mean of the non-textual representation given textual context, while the residual represents its deviation from that mean.

Text-Conditioned Diffusion Estimation. Directly modeling the conditional mapping from text to heterogeneous non-textual representations is difficult because $p(\mathbf{H}_m|\mathbf{T})$ may be highly multimodal. We therefore model this conditional distribution using diffusion. For $m \in \{a, v\}$, the conditional generative model is factorized as

$$p_{\theta_m}(\mathbf{x}_{0:K}|\mathbf{T}) = p(\mathbf{x}_K) \prod_{k=1}^{K} p_{\theta_m}(\mathbf{x}_{k-1}|\mathbf{x}_k, \mathbf{T}), \tag{2}$$

where $\mathbf{x}_0$ corresponds to the observed acoustic or visual representation and $p(\mathbf{x}_K) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. The diffusion model is trained to approximate the data conditional distribution, $p_{\theta_m}(\mathbf{H}_m|\mathbf{T}) \approx p_{\text{data}}(\mathbf{H}_m \mid \mathbf{T})$. Equivalently, this can be viewed as minimizing

$$\theta_m^\star = \arg\min_{\theta_m} \mathbb{E}_{\mathbf{T}} \left[D_{\text{KL}} \left(p_{\text{data}}(\mathbf{H}_m \mid \mathbf{T}) \,\|\, p_{\theta_m}(\mathbf{H}_m \mid \mathbf{T}) \right) \right]. \tag{3}$$

The standard denoising diffusion objective provides a tractable surrogate for this conditional likelihood objective and its derivation and reverse sampling procedure are given in Appendix F. Let $\widetilde{\mathbf{S}}_m$ denote the text-conditioned representation generated by the learned conditional diffusion model. We construct the corresponding residual as $\widetilde{\mathbf{P}}_m = \mathbf{H}_m - \widetilde{\mathbf{S}}_m$, such that $\mathbf{H}_m = \widetilde{\mathbf{S}}_m + \widetilde{\mathbf{P}}_m$. Importantly, this construction does not assume that $\widetilde{\mathbf{P}}_m$ is a semantically pure private representation. As detailed in Appendix F, the practical residual may contain text-unexplained modality-specific information, conditional distribution estimation error, and variation induced by diffusion sampling. Thus, its task relevance is determined by task-aware representation learning module introduced next.

3.3 Task-aware Shared-Private Representation Learning

Shared-Private Representation Construction. Let $\mathbf{u}_m, \mathbf{v}_m \in \mathbb{R}^d$ denote the shared and private reasoning inputs for modality $m \in \mathcal{M}$, and the detailed construction is given in Appendix G. Since the non-textual branches are conditioned on text, these representations are modality-associated rather than strictly unimodal. We map the two inputs into compact latent spaces $\mathbf{z}_m^{\text{sh}}, \mathbf{z}_m^{\text{pr}} \in \mathbb{R}^b$ with $b < d$. Let U_m and V_m denote the random variables associated with $\mathbf{u}_m$ and $\mathbf{v}_m$, respectively. The shared latent Z_m^{sh} is stochastic, whereas the private latent Z_m^{pr} is deterministic. Their joint reasoning representation is $Z_m = [Z_m^{\text{sh}}; Z_m^{\text{pr}}]$.

Asymmetric Information Bottleneck Formulation. Our objective is to compress information carried by the shared branch while requiring the joint shared-private representation to preserve sufficient information about the task target. We formulate this asymmetric information bottleneck as

$$\min_{\Theta_m} \quad I\big(U_m; Z_m^{\text{sh}}\big) \quad \text{s.t.} \quad I(Z_m; Y) \geq \kappa, \tag{4}$$

where Θ_m collects the parameters of the compact representation mapping and κ specifies the desired level of task information. With $\beta_{\text{IB}} > 0$ treated as a fixed trade-off coefficient, omitting the term $\beta_{\text{IB}}\kappa$, which is constant with respect to Θ_m, gives the Lagrangian-relaxed objective

$$\mathcal{J}_{\text{IB}}^m = I\big(U_m; Z_m^{\text{sh}}\big) - \beta_{\text{IB}} I(Z_m; Y), \tag{5}$$

where $\beta_{\text{IB}} > 0$. Note that this asymmetry is intentional, i.e., only the shared latent is subject to input-compression, whereas the private latent contributes to task relevance through the joint representation Z_m without receiving an additional information-compression penalty. This allows complementary modality-specific cues to remain available when they are useful for prediction. The stochastic shared encoder is modeled as $q_{\phi_m}(\mathbf{z}_m^{\text{sh}} \mid \mathbf{u}_m) = \mathcal{N}(\boldsymbol{\mu}_m, \text{diag}(\boldsymbol{\sigma}_m^2))$, with $\mathbf{z}_m^{\text{sh}} = \boldsymbol{\mu}_m + \boldsymbol{\sigma}_m \odot \boldsymbol{\epsilon}_m$ and $\boldsymbol{\epsilon}_m \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The private latent is obtained deterministically as $\mathbf{z}_m^{\text{pr}} = f_m^{\text{pr}}(\mathbf{v}_m)$. Using the standard Gaussian reference $r(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, the shared compression term admits the variational upper bound

$$I\big(U_m; Z_m^{\text{sh}}\big) \leq \mathbb{E}_{U_m} \left[D_{\text{KL}} \left(q_{\phi_m}(Z_m^{\text{sh}} \mid U_m) \,\|\, r(Z_m^{\text{sh}}) \right) \right]. \tag{6}$$

Conversely, the auxiliary task likelihood introduced in Sec. 3.4 is defined on the evidence representation E_m, which is a deterministic function of Z_m. By the data-processing inequality and a variational likelihood bound, it provides a tractable lower bound for preserving task information in Z_m.

Consequently, the KL compression loss and the auxiliary evidence loss form a variational surrogate of Eq. (5). The complete derivation is given in Appendix H. Thus, we use the dimension-normalized compression penalty

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{Mb} \sum_{m \in \mathcal{M}} \mathbb{E}_{U_m} \left[D_{\mathrm{KL}} \left(q_{\phi_m}(Z_m^{\mathrm{sh}} \mid U_m) \,\|\, \mathcal{N}(\mathbf{0}, \mathbf{I}) \right) \right]. \tag{7}$$

Task-Conditioned Residual Dependence. Noting that both branches may contain task-relevant information, unconditional decorrelation may suppress useful structure. Thus, we penalize only the residual dependence remaining after target adjustment. For a minibatch of size B, let $\mathbf{K}_m^{\mathrm{sh}}$, $\mathbf{K}_m^{\mathrm{pr}}$, and $\mathbf{K}_y \in \mathbb{R}^{B \times B}$ denote the centered RBF Gram matrices constructed from the shared posterior means $\boldsymbol{\mu}_m$, private compact latents $\mathbf{z}_m^{\mathrm{pr}}$, and task targets, respectively. We define the residual operator

$$\mathbf{R}_y = \mathbf{I}_B - \mathbf{K}_y \left(\mathbf{K}_y + \lambda_r \mathbf{I}_B \right)^{-1}, \tag{8}$$

where $\lambda_r > 0$ and residualize both branches using the same operator, i.e., $\widetilde{\mathbf{K}}_m^{\mathrm{sh}} = \mathbf{R}_y \mathbf{K}_m^{\mathrm{sh}} \mathbf{R}_y$, and $\widetilde{\mathbf{K}}_m^{\mathrm{pr}} = \mathbf{R}_y \mathbf{K}_m^{\mathrm{pr}} \mathbf{R}_y$. The remaining shared-private dependence is measured by the normalized residual-kernel coefficient

$$\rho_m = \frac{\left\langle \widetilde{\mathbf{K}}_m^{\mathrm{sh}}, \widetilde{\mathbf{K}}_m^{\mathrm{pr}} \right\rangle_F}{\left\| \widetilde{\mathbf{K}}_m^{\mathrm{sh}} \right\|_{F,\varepsilon} \left\| \widetilde{\mathbf{K}}_m^{\mathrm{pr}} \right\|_{F,\varepsilon}}, \tag{9}$$

yielding $\mathcal{L}_{\mathrm{red}} = (1/M) \sum_{m \in \mathcal{M}} \max(0, \rho_m)$. Thus, $\mathcal{L}_{\mathrm{red}}$ does not enforce unconditional independence between the two branches. Instead, it suppresses their remaining dependence after regularized adjustment for target-associated variation. The residual-kernel construction, numerical stabilization, and the properties of ρ_m are given in Appendix I.

3.4 Disagreement-Aware Latent State Reasoning

Task-Level Modality Evidence. For each modality $m \in \mathcal{M}$, the compact shared latent is first mapped into a shared space, denoted by $\bar{\mathbf{z}}_m^{\mathrm{sh}}$. The shared and private latents are then combined by a parameter-shared evidence encoder, i.e., $\mathbf{e}_m = g_{\mathrm{ev}} \left([\bar{\mathbf{z}}_m^{\mathrm{sh}}; \mathbf{z}_m^{\mathrm{pr}}] \right) \in \mathbb{R}^{d_e}$ A shared task likelihood $q_\psi(y \mid \mathbf{e}_m)$ maps all modality-associated evidence into the same target space. The corresponding auxiliary negative log-likelihood is

$$\mathcal{L}_{\mathrm{ev}} = -\frac{1}{M} \sum_{m \in \mathcal{M}} \mathbb{E} \left[\log q_\psi \left(Y \mid \mathbf{e}_m \right) \right]. \tag{10}$$

Shared task likelihood and supervision make modality evidence comparable in a common task space; its variational interpretation is derived in Appendix H.

Predictive Disagreement. Let $\boldsymbol{\xi}_m$ denote the task prediction induced by $\mathbf{e}_m$, represented as a scalar for regression and a probability vector for classification. We define pairwise predictive disagreement,

$$D_{mn} = \delta \left(\mathrm{sg}(\boldsymbol{\xi}_m), \mathrm{sg}(\boldsymbol{\xi}_n) \right), \tag{11}$$

where $D_{mm} = 0$ and sg denotes the stop-gradient and δ is a task-appropriate symmetric distance in the common output space. The task-specific instantiations are given in Appendix J.2. Importantly, disagreement alone is not sufficient to determine modality contributions. For scalar regression, consider $D_{mn} \propto |\widehat{y}_m - \widehat{y}_n|$. For any common translation c, we have $|(\widehat{y}_m + c) - (\widehat{y}_n + c)| = |\widehat{y}_m - \widehat{y}_n|$. Hence, pairwise disagreement is non-identifying with respect to the absolute task interpretation, given in Appendix J.3.

Latent Task-State Inference. Let $\mathcal{E} = \{\mathbf{e}_m\}_{m \in \mathcal{M}}$ denote the set of modality-associated evidence, $\mathbf{D} = [D_{mn}]_{m,n \in \mathcal{M}}$ the predictive disagreement structure, and $\mathbf{c}_{\mathrm{sh}}$ a shared-context representation inferred from $\{\bar{\mathbf{z}}_m^{\mathrm{sh}}\}_{m \in \mathcal{M}}$. We initialize a sample-specific latent task state from the shared context and update it once using modality evidence and predictive disagreement:

$$\mathbf{h}^{(0)} = f_{\mathrm{init}}(\mathbf{c}_{\mathrm{sh}}), \qquad \mathbf{h}^{(1)} = \Psi_{\phi,\vartheta} \left(\mathbf{h}^{(0)}, \mathcal{E}, \mathbf{D} \right). \tag{12}$$

Here, $\Psi_{\phi,\vartheta}$ denotes the state-refinement function, parameterized by the state-update parameters ϕ and routing parameters ϑ. The updated state $\mathbf{h}^{(1)}$ provides task context for assigning modality

contributions rather than estimating ground-truth modality reliability. For modality m, let $\mathbf{d}_m = [D_{mn}]_{n\in\mathcal{M}}$ denote its disagreement profile. Its state-conditioned contribution score is

$$a_m = s_\vartheta\left(\mathbf{h}^{(1)}, \mathbf{e}_m, \mathbf{d}_m\right), \tag{13}$$

where $s_\vartheta(\cdot)$ maps these inputs to an unnormalized scalar score. The normalized modality contributions are

$$w_m = \frac{\exp(a_m)}{\sum_{n\in\mathcal{M}} \exp(a_n)}, \qquad \sum_{m\in\mathcal{M}} w_m = 1. \tag{14}$$

Thus, modality evidence and predictive disagreement first refine the shared-context-initialized state, which then guides contribution assignment. Further details are provided in Appendix J.4.

State-Guided Fusion and Learning. The normalized contributions are converted into feature gates $g_m = Mw_m$, satisfying $\sum_{m\in\mathcal{M}} g_m = M$. Hence, uniform contributions yield $g_m = 1$, preserving the scale of the corresponding ungated fusion pathway. The compact reasoning branch determines modality contributions through latent-state refinement, while the full-width shared and private representations retained from the decomposition stage carry the information used for final prediction. Let $\mathbf{f}_{\text{sh}}$ and $\mathbf{f}_{\text{pr}}$ denote the gated shared-interaction and private fusion representations, respectively. They are integrated by stabilized projection-based residualization:

$$\mathbf{f} = \mathbf{f}_{\text{sh}} - \Pi^{\varepsilon}_{\mathbf{f}_{\text{pr}}}(\mathbf{f}_{\text{sh}}) + \mathbf{f}_{\text{pr}}, \tag{15}$$

where $\Pi^{\varepsilon}_{\mathbf{f}_{\text{pr}}}$ is defined in Appendix K. The fused representation is fed to task predictor, $\widehat{Y} = h_{\text{task}}(\mathbf{f})$. Further, the complete model is jointly optimized as

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{red}}\mathcal{L}_{\text{red}} + \lambda_{\text{ev}}\mathcal{L}_{\text{ev}}, \tag{16}$$

where the terms respectively optimize the final prediction, text-conditioned diffusion decomposition, shared-latent compression, task-conditioned residual dependence, and task-level evidence supervision. Detailed fusion and task-specific optimization are provided in Appendix K.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate InfoMind on four widely used multimodal benchmarks, including CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Zadeh et al., 2018), MIntRec (Zhang et al., 2022), and MIntRec2.0 (Zhang et al., 2024a). CMU-MOSI and CMU-MOSEI are standard multimodal sentiment analysis (MSA) benchmarks containing 2,199 and 22,856 utterance-level samples, respectively. MIntRec is a multimodal intent recognition (MIR) benchmark containing 2,224 samples across 20 fine-grained intent categories. MIntRec2.0 extends MIntRec to a substantially larger and more diverse setting, comprising 9,304 samples annotated with 30 fine-grained intent categories. Detailed dataset statistics and configurations are provided in Appendix B.

Evaluation Metrics. In this experiment, we adopt evaluation methods from previous work (Liang et al., 2021; Lv et al., 2021; Mao et al., 2022). For CMU-MOSI and CMU-MOSEI, we report Mean Absolute Error (MAE), Pearson correlation coefficient (Corr), 7-class accuracy (Acc-7), binary accuracy (Acc-2), and F1-score. For MIntRec and MIntRec2.0, we follow the standard MIR evaluation protocol and report accuracy (ACC), F1-score (F1), precision (P), recall (R), weighted F1-score (WF1), and weighted precision (WP). The detailed metric definitions are provided in Appendix C.

Implementation Details. Following recent approaches, we extract textual, acoustic, and visual representations using T5 (Raffel et al., 2020), Librosa (McFee et al., 2015), and OpenFace (Baltrusaitis et al., 2016), respectively. All experiments are conducted on a single NVIDIA A100 GPU. The implementation details and hyperparameter settings are provided in Appendix D.

Baselines. We compare InfoMind with state-of-the-art methods covering different multimodal learning paradigms. Details are provided in Appendix D. For MSA, we include semantic disentanglement methods (TCSP (Wu et al., 2021), DMD (Li et al., 2023), and MISA (Hazarika et al., 2020)), contrastive learning approaches (MCL-MCF (Fan et al., 2025) and HyCon (Mai et al., 2023)), cross-modal interaction methods (GLoMo (Zhuang et al., 2024), CRIL (Huang et al., 2023), and

Table 1: Main results on CMU-MOSI and CMU-MOSEI.

Models	CMU-MOSI					CMU-MOSEI				
	Acc7↑	Acc2↑	F1↑	Corr↑	MAE↓	Acc7↑	Acc2↑	F1↑	Corr↑	MAE↓
MISA$_{2020,MM}$	42.3	81.8/83.4	81.7/83.6	0.761	0.783	52.2	83.6/85.5	83.8/85.3	0.756	0.555
TCSP$_{2021,ACL}$	–	80.9/–	81.0/–	0.710	0.908	–	82.8/–	82.7/–	0.715	0.576
HyCon$_{2023,Trans.A.C.}$	46.6	–/85.2	–/85.1	0.790	0.713	52.8	–/85.4	–/85.6	0.776	0.601
AMML$_{2023,Trans.A.C.}$	46.3	–/84.9	–/84.8	0.723	0.769	52.4	–/85.3	–/85.2	0.614	0.776
DMD$_{2023,CVPR}$	45.6	–/84.5	–/84.4	–	–	54.5	–/85.0	–/84.9	–	–
CRIL$_{2023,MM}$	46.3	–/86.6	–/86.5	0.812	0.695	54.3	–/86.9	–/86.8	0.775	0.525
ALMT$_{2023,EMNLP}$	49.4	84.55/86.43	84.57/86.47	0.805	0.683	54.2	84.78/86.79	85.19/86.86	0.779	0.526
EMT$_{2024,Trans.A.C.}$	47.4	83.3/85.0	83.2/85.0	0.798	0.705	54.5	83.4/86.0	83.7/86.0	0.774	0.527
PEST$_{2024,KBS}$	–	–/86.1	–/86.1	0.796	0.723	–	–/85.3	–/85.1	0.761	0.542
KEBR$_{2024,MM}$	47.81	84.84/87.27	84.83/87.25	0.819	0.683	54.37	84.01/86.74	84.25/86.68	0.799	0.517
GLoMo$_{2024,MM}$	48.3	84.1/86.7	83.9/86.6	0.782	0.718	55.0	83.7/86.5	84.0/86.4	0.771	0.539
HyDiscGAN$_{2024,IJCAI}$	43.2	84.1/86.7	83.7/86.3	0.782	0.749	54.4	81.9/86.3	82.1/86.2	0.761	0.533
MCL-MCF$_{2025,Trans.A.C.}$	–	84.9/87.3	84.7/87.2	0.799	0.692	–	84.2/86.4	84.4/86.3	0.767	0.536
EMOE$_{2025,CVPR}$	46.5	-/84.76	-/84.78	–	0.719	53.55	-/85.77	-/85.75	–	0.527
DiffuFuse$_{2025,MM}$	50.7	86.1/87.8	85.9/87.5	0.791	0.672	55.4	86.7/88.1	86.4/87.7	0.776	0.521
InfoMind (Ours)	**52.04**	**87.45/88.87**	**87.46/88.89**	**0.887**	**0.561**	**55.87**	**89.9/89.2**	**89.8/88.9**	**0.837**	**0.496**

Table 2: Main results on MIntRec and MIntRec2.0.

Methods	MIntRec						MIntRec2.0					
	ACC↑	F1↑	P↑	R↑	WF1↑	WP↑	ACC↑	F1↑	P↑	R↑	WF1↑	WP↑
MISA$_{2020,MM}$	72.13	69.34	70.60	69.27	72.34	73.43	57.18	51.90	53.03	51.74	57.15	57.82
MulT$_{2019,ACL}$	71.69	67.71	68.78	67.76	71.38	71.79	58.58	51.64	54.74	51.49	57.43	57.49
MMIM$_{2021,EMNLP}$	71.73	68.92	69.51	69.21	71.60	72.21	56.35	50.48	52.52	50.85	55.44	56.44
MAG-BERT$_{2020,ACL}$	72.00	67.70	68.18	68.52	71.64	72.25	58.37	49.91	52.66	50.62	56.55	56.21
TCL-MAP$_{2024,AAAI}$	73.35	69.31	69.50	70.30	72.92	73.27	57.83	52.16	54.02	52.35	57.00	57.40
SDIF-DA$_{2024,ICASSP}$	71.28	68.08	69.32	67.82	70.98	71.32	57.93	51.97	53.38	52.04	57.23	57.71
MIntOOD$_{2025,TMM}$	72.81	69.70	70.91	69.49	72.62	73.06	57.92	51.56	56.91	51.29	56.75	58.67
MVCL-DAF$_{2025,AAAI}$	73.71	70.33	70.93	70.44	73.38	73.57	58.65	52.27	56.50	51.53	58.16	58.96
InfoMind (Ours)	**77.35**	**74.78**	**75.62**	**75.48**	**77.18**	**78.04**	**60.60**	**53.95**	**58.98**	**53.35**	**59.33**	**60.27**

PEST (Gan et al., 2024)), and recent multimodal fusion methods including EMT (Sun et al., 2024), AMML (Sun et al., 2023), ALMT (Zhang et al., 2023), KEBR (Zhu et al., 2024), HyDiscGAN (Wu et al., 2024), EMOE (Fang et al., 2025), and DiffuFuse (Lv et al., 2025). For MIR, we further compare with representative MSA-based and MIR-specific baselines, including MISA (Hazarika et al., 2020), MulT (Tsai et al., 2019), MMIM (Han et al., 2021), MAG-BERT (Rahman et al., 2020), TCL MAP (Zhou et al., 2024), SDIF-DA (Huang et al., 2024), MIntOOD (Zhang et al., 2025), and MVCL-DAF (Hu et al., 2025).

4.2 Results and Analysis

Multimodal Sentiment Analysis. Table 1 reports the results on CMU-MOSI and CMU-MOSEI. InfoMind achieves the strongest overall performance on both datasets. On CMU-MOSI, InfoMind achieves 52.04% Acc-7, 0.887 Corr, and 0.561 MAE. It also obtains the best Acc-2 and macro-F1 scores under both binary evaluation settings. On CMU-MOSEI, it reaches 55.87% Acc-7, 0.837 Corr, and 0.496 MAE, with the largest improvements appearing in Acc-2 and macro-F1. These results show that InfoMind consistently improves both discrete sentiment classification and continuous sentiment estimation on CMU-MOSI and CMU-MOSEI.

Multimodal Intent Recognition. Table 2 reports the results on MIntRec and MIntRec2.0. InfoMind obtains 77.35% ACC and 74.78% macro-F1 on MIntRec, and 60.60% ACC and 53.95% macro-F1 on MIntRec2.0. Together with the sentiment results, these findings extend the evidence for InfoMind from continuous sentiment estimation to fine-grained multiclass intent recognition. Since the main tables evaluate the complete model, we next examine which training objectives and reasoning components contribute to this performance.

Training-Objective Ablations. Table 3 evaluates the four auxiliary objectives. Removing $\mathcal{L}_{\text{ev}}$ causes the largest degradation: CMU-MOSI Acc-7 decreases from 52.04% to 45.04% and MAE

Table 3: Loss ablation results on CMU-MOSI and MIntRec. Higher is better for all metrics except MAE. Results on MIntRec are averaged over five random seeds.

Ablation	CMU-MOSI					MIntRec					
	Acc7↑	Acc2↑	F1↑	Corr↑	MAE↓	ACC↑	F1↑	P↑	R↑	WF1↑	WP↑
w/o $\mathcal{L}_{\text{diff}}$	49.71	87.50	87.51	0.876	0.582	76.76	74.51	75.01	75.38	76.73	77.79
w/o $\mathcal{L}_{\text{KL}}$	46.79	86.43	86.43	0.871	0.638	77.16	74.12	74.27	75.12	77.08	77.86
w/o $\mathcal{L}_{\text{red}}$	49.71	87.20	87.16	0.877	0.599	76.62	74.17	74.49	74.99	76.65	77.62
w/o $\mathcal{L}_{\text{ev}}$	45.04	85.37	85.35	0.875	0.651	75.67	72.95	74.12	73.71	75.63	77.01
InfoMind (Full)	**52.04**	**88.87**	**88.89**	**0.887**	**0.561**	**77.35**	**74.78**	**75.62**	**75.48**	**77.18**	**78.04**

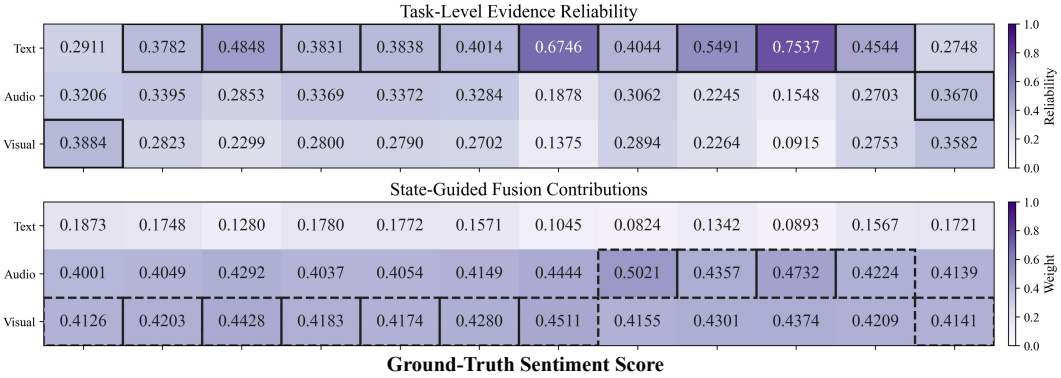


Figure 3: Evidence reliability (top) and state-guided fusion weights (bottom) for selected test samples ordered by ground-truth sentiment score. Solid boxes mark the modality with the highest evidence-reliability score in each sample column, whereas dashed boxes mark the modality receiving the largest fusion weight.

Table 4: Structural ablation of the disagreement-aware latent-state reasoning module on MIntRec and CMU-MOSI. Higher is better for all metrics except MAE.

Variant	MIntRec						CMU-MOSI				
	ACC↑	F1↑	P↑	R↑	WF1↑	WP↑	Acc7↑	Acc2↑	F1↑	Corr↑	MAE↓
w/o Disagreement Input	75.91	73.39	73.96	74.28	76.08	77.36	50.00	88.87	88.88	0.883	0.583
w/o State Refinement	76.40	73.79	74.71	74.23	76.34	77.18	51.60	88.72	88.70	0.881	0.575
Direct Gating	75.82	73.16	73.77	73.84	76.12	77.31	51.90	88.57	88.57	0.879	0.581
w/o Shared-Context Init.	75.60	72.33	72.73	73.34	75.57	76.60	50.44	88.41	88.42	0.883	0.568
InfoMind (Full)	**77.35**	**74.78**	**75.62**	**75.48**	**77.18**	**78.04**	**52.04**	**88.87**	**88.89**	**0.887**	**0.561**

increases from 0.561 to 0.651, while MIntRec ACC and macro-F1 decrease to 75.67% and 72.95%, respectively. Removing $\mathcal{L}_{\text{KL}}$ increases CMU-MOSI MAE to 0.638 and reduces Corr to 0.871. Removing $\mathcal{L}_{\text{diff}}$ or $\mathcal{L}_{\text{red}}$ also degrades overall performance. These results indicate complementary contributions from task-level evidence supervision, shared compression, diffusion decomposition, and dependence control. Additional visualizations are provided in Appendix E.

Structural Ablations. Table 4 evaluates the disagreement-aware latent-state reasoning module. *Direct Gating* maps the same context, evidence, and disagreement inputs directly to modality weights, without constructing or refining an intermediate task state. Compared with this variant, the full model improves MIntRec ACC and macro-F1 by 1.26 and 1.36 percentage points and reduces CMU-MOSI MAE from 0.581 to 0.561. Removing state refinement lowers MIntRec ACC to 76.40% and increases CMU-MOSI MAE to 0.575. Removing disagreement raises MAE to 0.583, although binary accuracy remains unchanged, while removing shared-context initialization produces the lowest MIntRec ACC and macro-F1. Together, these comparisons support one-step, context-initialized state refinement over direct gating, although the gains vary across tasks and metrics.

Evidence Reliability and Disagreement Analysis. Figure 3 compares how closely each modality-associated prediction matches the ground-truth sentiment with the modality contributions learned by InfoMind for selected CMU-MOSI samples. For visualization only, reliability is derived from the distance between each modality-associated prediction and the ground-truth sentiment and is not



Figure 4: InfoMind predicts the correct intent in all three displayed cases. In Sample 311, its prediction agrees with the text branch, which predicts Praise, while both the audio and visual branches predict Inform.

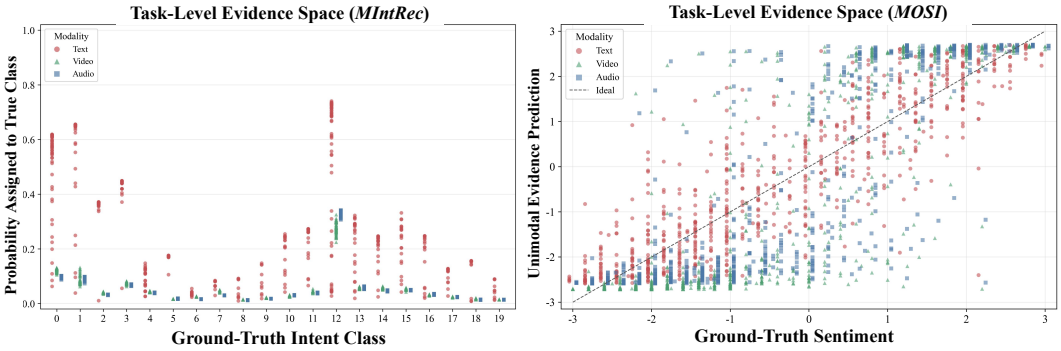


Figure 5: Modality-associated evidence predictions. Left: probabilities assigned to the ground-truth intent class. Right: auxiliary sentiment predictions versus ground-truth scores; the dashed line denotes ideal predictions ($\hat{y} = y$). The plots reveal differences in auxiliary prediction behavior across modalities and tasks.

used as an input to the fusion module. In the displayed samples, reliability and learned contributions often produce different modality rankings, showing that the contributions are not direct copies of individual predictive reliability.

Figure 4 presents three MIntRec examples with conflicting modality-associated predictions. InfoMind predicts the correct intent in all three cases. In Sample 311, the final prediction agrees with the text branch, which differs from the agreeing audio and visual branches. This case shows that the final decision does not necessarily follow a simple majority vote across modality-associated predictions.

Task-Level Evidence Predictions. Figure 5 shows the branch-level auxiliary predictions used to compute disagreement before final fusion. On MIntRec, confidence in the ground-truth class varies across modalities and intents. On CMU-MOSI, the distance from the identity line measures auxiliary prediction error. The variation in ground-truth-class confidence and prediction error shows that evidence informativeness depends on both the modality and the task.

5 Conclusion

We introduced **InfoMind**, which interprets task-level evidence and predictive disagreement through a sample-specific latent state before assigning modality contributions. Initialized from shared context, the state is updated once to guide fusion of full-width shared and private representations. Evaluations on four benchmarks spanning continuous sentiment estimation and multiclass intent recognition show consistent gains, supporting the effectiveness and cross-task applicability of InfoMind. Ablations further support evidence supervision and state refinement, with refinement outperforming direct gating. These findings support separating evidence interpretation from contribution assignment when multimodal cues conflict.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, drafting and revising parts of the manuscript, literature retrieval, and discussions of research methodology, experimental design, and result interpretation. All AI-assisted content was independently reviewed, verified, and revised by the authors. The authors take full responsibility for the final content of the paper, including its technical claims, experimental results, and conclusions.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed model architecture, training objectives, and optimization procedure in the main text. Dataset preprocessing, implementation details, hyperparameter settings, and additional experimental results are provided in the appendix. These details are intended to facilitate reproduction of the reported results.

REFERENCES

Tadas Baltrusaitis, Peter Robinson, and Louis-Philippe Morency. OpenFace: An open source facial behavior analysis toolkit. In *2016 IEEE Winter Conference on Applications of Computer Vision*, pp. 1–10, 2016.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5608–5628, 2024.

Henry S. Cheang and Marc D. Pell. The sound of sarcasm. *Speech Communication*, 50(5):366–381, 2008. doi: 10.1016/j.specom.2007.11.003.

Quanqi Du, Loic de Langhe, Els Lefever, and Veronique Hoste. LDW: Label divergence weighting for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland*, pp. 12342–12351, 2025.

Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002.

Cunhang Fan, Kang Zhu, Jianhua Tao, Guofeng Yi, Jun Xue, and Zhao Lv. Multi-level contrastive learning: Hierarchical alleviation of heterogeneity in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 16(1):207–222, 2025.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. EMOE: Modality-specific enhanced dynamic emotion experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14314–14324, 2025.

Zihan Fang, Zheng Lin, Senkang Hu, Yihang Tao, Yiqin Deng, Xianhao Chen, and Yuguang Fang. Dynamic uncertainty-aware multimodal fusion for outdoor health monitoring. *IEEE Transactions on Mobile Computing*, 25(10):15863–15878, 2026.

Chenquan Gan, Yu Tang, Xiang Fu, Qingyi Zhu, Deepak Kumar Jain, and Salvador García. Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation. *Knowledge-Based Systems*, 299:111982, 2024.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, 2021.

Zhiyuan Han, Beier Zhu, Yanlong Xu, Peipei Song, and Xun Yang. Benchmarking and bridging emotion conflicts for multimodal emotion reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland*, pp. 5528–5537, 2025.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1122–1131, 2020.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30183–30193, 2026.

Bo Hu, Kai Zhang, Yanghai Zhang, and Yuyang Ye. Adaptive multimodal fusion: Dynamic attention allocation for intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 17267–17275, 2025.

Jian Huang, Yanli Ji, Yang Yang, and Heng Tao Shen. Cross-modality representation interactive learning for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 426–434, 2023.

Shijue Huang, Libo Qin, Bingbing Wang, Geng Tu, and Ruifeng Xu. Sdif-da: A shallow-to-deep interaction framework with data augmentation for multi-modal intent detection. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 10206–10210, 2024.

Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning? (Provably). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9226–9259, 2022.

Haoyu Jiang, Xiaoliang Chen, Duoqian Miao, Xiaolin Qin, Xianyong Li, and Yajun Du. CICA: Coupling confidence-aware pretraining with confidence-informed attention for robust multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37842–37851, 2026.

Peiyuan Jiang, Yao Liu, Qiao Liu, Zongshun Zhang, Jiaye Yang, Lu Liu, and Daibing Yao. DRKF: Decoupled representations with knowledge fusion for multimodal emotion recognition. In *Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland*, pp. 955–964, 2025.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16049–16096, 2023.

Konrad P. Körding, Ulrik Beierholm, Wei Ji Ma, Steven Quartz, Joshua B. Tenenbaum, and Ladan Shams. Causal inference in multisensory perception. *PLoS ONE*, 2(9):e943, 2007. doi: 10.1371/journal.pone.0000943.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6631–6640, 2023.

Tao Liang, Guosheng Lin, Lei Feng, Yan Zhang, and Fengmao Lv. Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8128–8136, 2021.

Kejun Liu, Zhe Chen, Yuanyuan Liu, Ke Wang, Yibing Zhan, Wei Xiang, and Hongyan Zhang. World-model inspired emotion-aware token refinement for training-free multimodal emotion recognition. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2554–2562, 2021.

Xiongjian Lv, Yimin Wen, and Hang Yu. DiffuFuse: Diffusion-driven dual-stream fusion framework for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia, Dublin, Ireland*, pp. 8458–8467, 2025.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14 (3):2276–2289, 2023.

Huisheng Mao, Ziqi Yuan, Hua Xu, Wenmeng Yu, Yihe Liu, and Kai Gao. M-SENA: An integrated platform for multimodal sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Dublin, Ireland*, pp. 204–213, 2022.

Brian McFee, Colin Raffel, Dawen Liang, Daniel P. W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th Python in Science Conference*, pp. 18–25, 2015.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369, 2020.

Ladan Shams and Ulrik R. Beierholm. Causal inference in perception. *Trends in Cognitive Sciences*, 14(9):425–432, 2010. doi: 10.1016/j.tics.2010.07.001.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2024.

Ya Sun, Sijie Mai, and Haifeng Hu. Learning to learn better unimodal representations via adaptive multimodal meta-learning. *IEEE Transactions on Affective Computing*, 14(3):2209–2223, 2023.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019.

Yan Wang, Fuyuan Cao, and Xingwang Zhao. Conflict-aware adaptive cross-reconstruction for multimodal sentiment analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15880–15889, 2026.

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67584–67622, 2025.

Yang Wu, Zijie Lin, Yanyan Zhao, Bing Qin, and Li-Nan Zhu. A text-centered shared-private framework via cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4730–4738, 2021.

Zhuojia Wu, Qi Zhang, Duoqian Miao, Kun Yi, Wei Fan, and Liang Hu. HyDiscGAN: A hybrid distributed cGAN for audio-visual privacy preservation in multimodal sentiment analysis. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Zequn Yang, Hongfa Wang, and Di Hu. Efficient quantification of multimodal interaction at sample level. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 71302–71317, 2025.

Zequn Yang, Yake Wei, Haotian Ni, Zhihao Xu, and Di Hu. Information-theoretic decomposition for multimodal interaction learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30278–30287, 2026.

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

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2236–2246, 2018.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. MIntRec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1688–1697, 2022.

Hanlei Zhang, Xin Wang, Hua Xu, Qianrui Zhou, Kai Gao, Jianhua Su, Jinyue Zhao, Wenrui Li, and Yanting Chen. MIntRec2.0: A large-scale benchmark dataset for multimodal intent recognition and out-of-scope detection in conversations. In *The Twelfth International Conference on Learning Representations*, 2024a.

Hanlei Zhang, Qianrui Zhou, Hua Xu, Jianhua Su, Roberto Evans, and Kai Gao. Multimodal classification and out-of-distribution detection for multimodal intent understanding. *IEEE Transactions on Multimedia*, 27:9887–9901, 2025.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, 2023.

Yedi Zhang, Peter E. Latham, and Andrew M. Saxe. Understanding unimodal bias in multimodal deep linear networks. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59100–59125, 2024b.

Qianrui Zhou, Hua Xu, Hao Li, Hanlei Zhang, Xiaohan Zhang, Yifan Wang, and Kai Gao. Token-level contrastive learning with modality-aware prompting for multimodal intent recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17114–17122, 2024.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Fuji Ren. KEBR: Knowledge enhanced self-supervised balanced representation for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 5732–5741, 2024.

Yan Zhuang, Yanru Zhang, Zheng Hu, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. GLoMo: Global-local modal fusion for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1800–1809, 2024.

Shihao Zou and Wei Wei. CoRiM: Conflict-driven risk minimization for dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37821–37830, 2026.

A Overview

The appendix provides additional experimental and implementation details, theoretical derivations, and qualitative analyses of InfoMind to complement the main paper. Specifically:

- **Appendix B** provides detailed descriptions and statistics of the CMU-MOSI, CMU-MOSEI, MIntRec, and MIntRec2.0 datasets.
- **Appendix C** provides the definitions and evaluation protocols for the metrics used in multimodal sentiment analysis and multimodal intent recognition.
- **Appendix D** provides implementation details, hyperparameter settings, and detailed descriptions of the baselines used for comparison.
- **Appendix E** provides additional representation visualizations for qualitatively analyzing the effects of the proposed training objectives.
- **Appendix F** provides the complete formulation of the text-conditioned dual-stream diffusion module, including the forward diffusion process, denoising objective, and reverse-process formulation.
- **Appendix G** details the construction of the shared and private reasoning inputs from the sequence-level multimodal representations.
- **Appendix H** provides the variational derivation of the asymmetric information-bottleneck formulation used for task-aware shared-private representation learning.
- **Appendix I** provides the detailed construction and properties of the task-conditioned residual-dependence regularizer.
- **Appendix J** provides details of task-level evidence construction, predictive disagreement, latent-state refinement, and the interpretation of modality contribution assignment.
- **Appendix K** details how state-derived modality gates integrate the full-width shared and private representations retained from the text-conditioned decomposition stage and summarizes the overall learning objective.

B Dataset Details

- **CMU-MOSI.** CMU-MOSI (Zadeh et al., 2016) is a widely used multimodal sentiment analysis benchmark consisting of 2,199 utterance-level video segments. Following the standard split, we use 1,284 samples for training, 229 for validation, and 686 for testing. The acoustic and visual features are sampled at 12.5 Hz and 15 Hz, respectively. Each utterance is annotated with a sentiment score ranging from -3 to 3, where sign indicates polarity and absolute value reflects sentiment intensity.
- **CMU-MOSEI.** CMU-MOSEI (Zadeh et al., 2018) extends CMU-MOSI with a larger and more diverse collection of multimodal utterances. Following the standard preprocessing protocol, we use 22,856 YouTube movie review clips, including 16,326 training samples, 1,871 validation samples, and 4,659 test samples. The sentiment annotations follow the same scale from -3 to 3. The acoustic and visual features are sampled at 20 Hz and 15 Hz, respectively.
- **MIntRec.** MIntRec (Zhang et al., 2022) is a benchmark designed for multimodal intent recognition. It contains 2,224 samples with 20 fine-grained intent categories. Following the official split, we use 1,334 samples for training, 445 for validation, and 445 for testing. It comprises 11 intent categories related to emotions or attitudes and 9 related to goal achievement.
- **MIntRec2.0.** MIntRec2.0 (Zhang et al., 2024a) extends MIntRec to a larger and more diverse multimodal intent recognition setting. It contains 9,304 samples annotated with 30 fine-grained intent categories. Following the official split, we use 6,165 samples for training, 1,106 for validation, and 2,033 for testing. Compared with MIntRec, MIntRec2.0 substantially increases both the data scale and the diversity of intent categories, providing a more challenging setting for evaluating fine-grained multimodal intent recognition.

Table 5: Main hyperparameters of InfoMind.

Setting	CMU-MOSEI	CMU-MOSI	MIntRec	MIntRec2.0
Batch size	64	64	16	16
Learning rate	1×10^{-5}	1×10^{-5}	1×10^{-5}	7×10^{-6}
Epoch number	50	50	40	40
Dropout	0.2	0.2	0.2	0.2
Patience	20	20	5	5
Diffusion steps	50	50	50	50
State dimension	128	256	256	256
IB bottleneck dimension	128	128	128	128

Note that these datasets provide complementary evaluation scenarios, i.e., CMU-MOSI and CMU-MOSEI examine multimodal sentiment understanding at different data scales, while MIntRec and MIntRec2.0 evaluate the generality of InfoMind on fine-grained multimodal intent recognition under different dataset scales and category complexities.

C Evaluation Details

- For CMU-MOSI and CMU-MOSEI, we report MAE and Corr. MAE measures the average absolute difference between predicted and ground-truth sentiment scores. Corr measures the Pearson correlation between predictions and labels, reflecting the consistency of sentiment intensity estimation. For classification evaluation, Acc-2 and F1 follow two commonly used settings. The *All* setting performs negative/non-negative classification while retaining samples with zero sentiment scores. The *Non0* setting removes zero-valued samples and performs negative/positive classification.
- For MIntRec and MIntRec2.0, we report accuracy, macro-averaged F1-score, precision, and recall, together with weighted F1-score and weighted precision. The macro-averaged metrics assign equal importance to each intent category, whereas the weighted metrics account for class-frequency differences.

D Implementation Details and Baselines

For all datasets, the extracted multimodal representations are fed into InfoMind for shared-private decomposition and disagreement-aware reasoning. The learning rate is adjusted according to validation performance. Dropout is applied after orthogonal fusion for regularization. The detailed hyperparameter configurations are summarized in Table 5.

Multimodal Sentiment Analysis Baselines.

- **MISA** (Hazarika et al., 2020): MISA projects each modality into modality-invariant and modality-specific subspaces to capture cross-modal commonalities and modality-specific characteristics, respectively. The resulting representations are jointly used for multimodal fusion and sentiment prediction.
- **TCSP** (Wu et al., 2021): TCSP adopts a text-centered framework that uses cross-modal prediction to distinguish shared and private semantics in acoustic and visual inputs. It integrates these semantics with textual features, using shared information to reinforce textual semantics and private information to provide complementary cues.
- **DMD** (Li et al., 2023): DMD decomposes each modality's representation into modality-irrelevant and modality-exclusive components. It applies a separate graph distillation unit to each component, dynamically learning cross-modal knowledge transfer weights to improve the discriminative representations of individual modalities.
- **HyCon** (Mai et al., 2023): HyCon combines intra-modal, inter-modal, and semi-contrastive learning to capture cross-modal interactions and relationships between classes while re-

ducing modality gaps. An additional refinement term regularizes the contrastive learning process.

- **MCL-MCF** (Fan et al., 2025): MCL-MCF combines multi-level contrastive learning with multi-layer convolutional fusion. Its hierarchical objectives address heterogeneity among unimodal representations, between unimodal and fused representations, and among higher-level fused features, while progressively integrating shallow and deep features.
- **GLoMo** (Zhuang et al., 2024): GLoMo employs modality-specific mixture-of-experts layers to select and aggregate informative local representations. A global-guided fusion module subsequently combines local and global representations to retain information at both levels of granularity.
- **CRIL** (Huang et al., 2023): CRIL uses textual guidance to learn compact semantic tokens for acoustic and visual modalities and aligns their semantic relationships by enforcing consistency among modality-specific self-attention matrices. A two-stage interactive fusion mechanism then integrates the aligned representations for sentiment prediction.
- **PEST** (Gan et al., 2024): PEST translates textual, acoustic, and visual features into a common space to reduce modality heterogeneity and establish initial cross-modal interactions. A dynamical propagation model further refines these interactions to capture emotional correlations across modalities.
- **EMT** (Sun et al., 2024): EMT models efficient cross-modal interactions between utterance-level multimodal context and local unimodal features. Its complete framework, EMT-DLFR, additionally combines low-level feature reconstruction with high-level representation alignment between complete and incomplete views to improve robustness to missing modality features.
- **AMML** (Sun et al., 2023): AMML meta-trains unimodal networks through modality-specific optimization with adaptive learning rates and adapts the resulting representations for fusion through meta-testing on multimodal tasks. A distribution transformation layer further regularizes unimodal feature distributions to reduce cross-modal distributional discrepancies.
- **ALMT** (Zhang et al., 2023): ALMT uses language features at multiple scales to guide the construction of an adaptive hyper-modality representation from acoustic and visual inputs. This representation suppresses sentiment-irrelevant and conflicting information before multimodal fusion.
- **KEBR** (Zhu et al., 2024): KEBR performs knowledge-enhanced pretraining on unlabeled videos through sentiment word masking and sentiment intensity prediction. It combines text-based cross-modal fusion with a cosine-based constraint to incorporate nonverbal cues while balancing multimodal information during joint learning.
- **HyDiscGAN** (Wu et al., 2024): HyDiscGAN addresses audio-visual privacy preservation through a hybrid distributed conditional GAN framework. It generates synthetic acoustic and visual features conditioned on shareable, de-identified text, enabling multimodal sentiment prediction while protecting the corresponding real modalities.
- **EMOE** (Fang et al., 2025): EMOE combines a mixture of modality experts with unimodal distillation. The expert mechanism learns sample-dependent modality weights, while distillation preserves unimodal predictive capabilities within the fused representation.
- **DiffuFuse** (Lv et al., 2025): DiffuFuse uses text-conditioned diffusion denoising to predict shared semantics and adaptively distinguish shared and private information in non-textual features. A dual-stream fusion mechanism then models interactions across modalities and between the two semantic types.

Multimodal Intent Recognition Baselines.

- **MISA** (Hazarika et al., 2020): MISA projects each modality into modality-invariant and modality-specific subspaces to capture cross-modal commonalities and distinctive modality characteristics, respectively. These representations are jointly used for multimodal fusion and downstream prediction.

- **MulT** (Tsai et al., 2019): MulT uses directional pairwise cross-modal attention to model interactions between multimodal sequences at different time steps. It adapts representations from one modality to another, capturing cross-modal dependencies without requiring explicit temporal alignment.

- **MMIM** (Han et al., 2021): MMIM hierarchically maximizes mutual information both between pairs of unimodal representations and between the fused representation and individual unimodal inputs. These objectives are optimized jointly with the downstream task to retain task-relevant information during fusion.

- **MAG-BERT** (Rahman et al., 2020): MAG-BERT augments pretrained BERT with a multimodal adaptation gate that incorporates acoustic and visual cues during fine-tuning. The gate generates a nonverbal-conditioned shift in BERT's internal representations, enabling the language model to process multimodal inputs.

- **TCL-MAP** (Zhou et al., 2024): TCL-MAP constructs modality-aware prompts through similarity-based alignment and cross-modal attention. It forms augmented samples using these prompts and intent labels, then applies an NT-Xent objective to label-token representations to guide nonverbal semantic learning with intent-label information.

- **SDIF-DA** (Huang et al., 2024): SDIF-DA progressively aligns and fuses textual, acoustic, and visual features through a shallow-to-deep interaction module. It additionally uses ChatGPT-based data augmentation to address the scarcity of labeled training data and transfer knowledge from the language model.

- **MIntOOD** (Zhang et al., 2025): MIntOOD combines dynamically weighted multimodal fusion with representation learning for in-distribution intent classification and out-of-distribution detection. It synthesizes pseudo-OOD samples from convex combinations of in-distribution data and couples coarse-grained ID/OOD discrimination with fine-grained, confidence-guided intent classification and instance-level contrastive learning.

- **MVCL-DAF** (Hu et al., 2025): MVCL-DAF dynamically allocates attention both within and across modalities according to each input sample. It combines independent unimodal intent prediction with multimodal fusion and uses text-guided multi-view contrastive learning to coordinate the modality-specific representations.

E Representation Visualization

Figure 6 compares t-SNE projections of the raw multimodal features and representations learned with different training objectives on CMU-MOSI. The full model exhibits a visually more ordered sentiment gradient, while the ablated variants show different patterns of local mixing. These visualizations provide qualitative context for the training-objective ablations. Since t-SNE layouts depend on initialization and hyperparameters, their geometry alone does not establish representation quality or explain the effect of state refinement.

F Details of Dual-Stream Diffusion

We provide the complete probabilistic formulation, training objective, and reverse sampling procedure of the dual-stream diffusion module introduced in Sec. 3.2. The textual-visual and textual-acoustic streams follow the same conditional diffusion formulation with modality-specific parameters. And following the notation in Sec. 3.2, for each $m \in \{a, v\}$, $\mathbf{H}_m$ denotes the corresponding non-textual representation. For each aligned position $\ell \in \{1, \ldots, L\}$, let $\mathbf{T}^\ell, \mathbf{H}_m^\ell \in \mathbb{R}^d$ denote the corresponding textual and non-textual feature vectors. The diffusion process is formulated position-wise with $\mathbf{x}_0 = \mathbf{H}_m^\ell$ conditioned on $\mathbf{T}^\ell$. The sequence-level notation $p_{\theta_m}(\mathbf{H}_m \mid \mathbf{T})$ used in the main text serves as a compact representation of these aligned position-wise conditional processes. For clarity, we suppress the position index ℓ when no ambiguity arises.

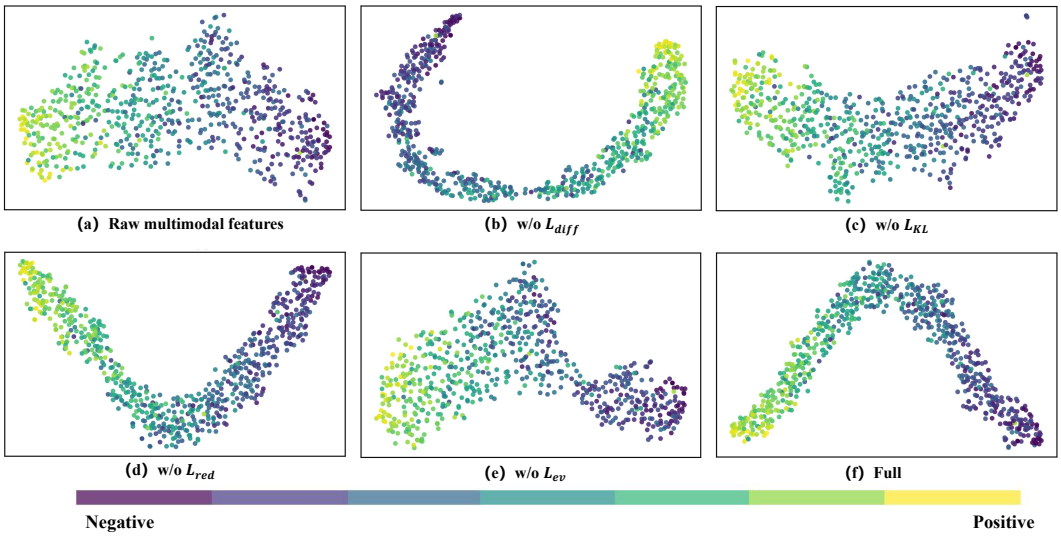


Figure 6: t-SNE visualization of multimodal representations on CMU-MOSI under different training-objective settings. Each point is colored according to its continuous sentiment score, ranging from negative (purple) to positive (yellow). The panels compare the raw multimodal features and variants without $\mathcal{L}_{\text{diff}}$, $\mathcal{L}_{\text{KL}}$, $\mathcal{L}_{\text{red}}$, or $\mathcal{L}_{\text{ev}}$, together with the full model. The full model exhibits a clearer organization along the sentiment continuum. Since t-SNE is sensitive to initialization and hyperparameters, the figure is intended for qualitative interpretation and complements, rather than replaces, the quantitative ablation results.

F.1 Forward Diffusion Process

For a clean non-textual feature $\mathbf{x}_0$, the forward diffusion process defines the Markov chain

$$q(\mathbf{x}_{1:K} \mid \mathbf{x}_0) = \prod_{k=1}^{K} q(\mathbf{x}_k \mid \mathbf{x}_{k-1}), \tag{17}$$

where each transition is given by

$$q(\mathbf{x}_k \mid \mathbf{x}_{k-1}) = \mathcal{N}\left(\sqrt{\alpha_k}\mathbf{x}_{k-1}, (1-\alpha_k)\mathbf{I}\right), \tag{18}$$

with $\alpha_k = 1 - \beta_k$ and $\beta_k \in (0,1)$ denoting the predefined noise schedule. Equivalently, the forward transition can be written as

$$\mathbf{x}_k = \sqrt{\alpha_k}\mathbf{x}_{k-1} + \sqrt{1-\alpha_k}\boldsymbol{\epsilon}_k, \tag{19}$$

where $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the Gaussian perturbations are independently sampled across timesteps. Defining the cumulative signal coefficient as

$$\bar{\alpha}_k = \prod_{s=1}^{k} \alpha_s, \tag{20}$$

the marginal distribution at an arbitrary diffusion timestep admits the closed form

$$q(\mathbf{x}_k \mid \mathbf{x}_0) = \mathcal{N}\left(\sqrt{\bar{\alpha}_k}\mathbf{x}_0, (1-\bar{\alpha}_k)\mathbf{I}\right). \tag{21}$$

Thus, $\mathbf{x}_k$ can be sampled directly from $\mathbf{x}_0$ as

$$\mathbf{x}_k = \sqrt{\bar{\alpha}_k}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_k}\boldsymbol{\epsilon}, \tag{22}$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The variable $\boldsymbol{\epsilon}$ in Eq. (22) represents the aggregate Gaussian perturbation at timestep k, whereas $\boldsymbol{\epsilon}_k$ in Eq. (19) denotes the noise introduced at a single forward step.

F.2 Text-Conditioned Reverse Diffusion

The reverse process reconstructs the non-textual representation under textual conditioning. For modality $m \in \{a, v\}$, the conditional reverse distribution is factorized as

$$p_{\theta_m}\left(\mathbf{x}_{0:K} \mid \mathbf{T}^{\ell}\right) = p(\mathbf{x}_K) \prod_{k=1}^{K} p_{\theta_m}\left(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{T}^{\ell}\right), \tag{23}$$

where $p(\mathbf{x}_K) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. Each reverse transition is modeled by

$$p_{\theta_m}\left(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{T}^\ell\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\theta_m}(\mathbf{x}_k, \mathbf{T}^\ell, k), \beta_k \mathbf{I}\right), \tag{24}$$

where $\boldsymbol{\mu}_{\theta_m}$ denotes the text-conditioned reverse mean and β_k denotes the reverse variance. Under the standard noise-prediction parameterization,

$$\boldsymbol{\mu}_{\theta_m}\left(\mathbf{x}_k, \mathbf{T}^\ell, k\right) = \frac{1}{\sqrt{\alpha_k}}\left(\mathbf{x}_k - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\boldsymbol{\epsilon}_{\theta_m}\left(\mathbf{x}_k, \mathbf{T}^\ell, k\right)\right), \tag{25}$$

where $\boldsymbol{\epsilon}_{\theta_m}$ estimates the perturbation conditioned jointly on the current noisy non-textual representation and its aligned textual context. The textual representation therefore acts as a conditioning variable throughout the reverse trajectory. The acoustic and visual streams share the same probabilistic formulation while using separate parameter sets θ_a and θ_v.

F.3 Conditional Diffusion Objective

For modality $m \in \{a, v\}$, the conditional diffusion model seeks to learn $p_{\theta_m}(\mathbf{x}_0 \mid \mathbf{T}^\ell)$. The corresponding conditional negative log-likelihood is

$$\mathcal{L}^m_{\mathrm{NLL}} = \mathbb{E}_{p_{\mathrm{data}}}\left[-\log p_{\theta_m}\left(\mathbf{x}_0 \mid \mathbf{T}^\ell\right)\right]. \tag{26}$$

Since the intermediate diffusion states are latent, the forward process $q(\mathbf{x}_{1:K} \mid \mathbf{x}_0)$ is used as a variational distribution. Applying Jensen's inequality gives

$$-\log p_{\theta_m}\left(\mathbf{x}_0 \mid \mathbf{T}^\ell\right) \le \mathbb{E}_q\left[\log \frac{q(\mathbf{x}_{1:K} \mid \mathbf{x}_0)}{p_{\theta_m}\left(\mathbf{x}_{0:K} \mid \mathbf{T}^\ell\right)}\right] := \mathcal{L}^m_{\mathrm{VLB}}. \tag{27}$$

Using Markov factorizations of the forward and reverse processes, this bound can be written as

$$\mathcal{L}^m_{\mathrm{VLB}} = \mathbb{E}_q\Bigg[D_{\mathrm{KL}}\left(q(\mathbf{x}_K \mid \mathbf{x}_0) \,\|\, p(\mathbf{x}_K)\right) + \sum_{k=2}^{K} D_{\mathrm{KL}}\left(q(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{x}_0) \,\|\, p_{\theta_m}(\mathbf{x}_{k-1} \mid \mathbf{x}_k, \mathbf{T}^\ell)\right) - \log p_{\theta_m}\left(\mathbf{x}_0 \mid \mathbf{x}_1, \mathbf{T}^\ell\right)\Bigg]. \tag{28}$$

Under the Gaussian noise-prediction parameterization, the timestep-dependent terms reduce, up to known weights and constants, to the standard denoising objective

$$\mathcal{L}^m_{\mathrm{diff}} = \mathbb{E}_{\mathbf{x}_0, \mathbf{T}^\ell, k, \boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_{\theta_m}\left(\mathbf{x}_k, \mathbf{T}^\ell, k\right)\right\|_2^2\right], \tag{29}$$

where $m \in \{a, v\}$ and k is sampled from $\{1, \ldots, K\}$ and $\mathbf{x}_k$ is obtained from Eq. (22). The total dual-stream diffusion objective is

$$\mathcal{L}_{\mathrm{diff}} = \mathcal{L}^v_{\mathrm{diff}} + \mathcal{L}^a_{\mathrm{diff}}, \tag{30}$$

where the two modality-specific diffusion objectives contribute equally.

F.4 Reverse Sampling

After training, the learned reverse process can be interpreted as starting from $\widehat{\mathbf{x}}_K \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and proceeds recursively from $k = K$ to $k = 1$. A reverse sampling step is written as

$$\widehat{\mathbf{x}}_{k-1} = \frac{1}{\sqrt{\alpha_k}}\left(\widehat{\mathbf{x}}_k - \frac{\beta_k}{\sqrt{1-\bar{\alpha}_k}}\boldsymbol{\epsilon}_{\theta_m}\left(\widehat{\mathbf{x}}_k, \mathbf{T}^\ell, k\right)\right) + \sqrt{\beta_k}\mathbf{z}_k, \tag{31}$$

where $\mathbf{z}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $k > 1$, and $\mathbf{z}_1 = \mathbf{0}$. We distinguish the reverse-chain variable $\widehat{\mathbf{x}}_k$ from the forward-corrupted variable $\mathbf{x}_k$. The latter is constructed by perturbing an observed non-textual representation during training, whereas the former belongs to a conditional generative trajectory initialized from Gaussian noise.

F.5 Shared-Private Representation Construction

The learned reverse process defines the conditional distribution $p_{\theta_m}(\mathbf{H}_m \mid \mathbf{T})$, where $m \in \{a, v\}$. The corresponding model conditional mean is

$$\widehat{\mathbf{S}}_m(\mathbf{T}) = \mathbb{E}_{p_{\theta_m}}[\mathbf{H}_m \mid \mathbf{T}]. \tag{32}$$

A reverse trajectory provides a conditional realization $\widetilde{\mathbf{S}}_m \sim p_{\theta_m}(\mathbf{H}_m \mid \mathbf{T})$, which is used as the generated shared representation. The corresponding residual representation is

$$\widetilde{\mathbf{P}}_m = \mathbf{H}_m - \widetilde{\mathbf{S}}_m. \tag{33}$$

Therefore, $\mathbf{H}_m = \widetilde{\mathbf{S}}_m + \widetilde{\mathbf{P}}_m$ holds by construction. To clarify the interpretation of the residual, define the ideal text-predictable component as

$$\mathbf{S}_m^\star = \mathbb{E}_{p_{\text{data}}}[\mathbf{H}_m \mid \mathbf{T}]. \tag{34}$$

The practical residual can then be decomposed as

$$\widetilde{\mathbf{P}}_m = \mathbf{H}_m - \widetilde{\mathbf{S}}_m = \underbrace{\mathbf{H}_m - \mathbf{S}_m^\star}_{\text{text-unexplained component}} + \underbrace{\mathbf{S}_m^\star - \widehat{\mathbf{S}}_m}_{\text{distribution-estimation discrepancy}} + \underbrace{\widehat{\mathbf{S}}_m - \widetilde{\mathbf{S}}_m}_{\text{sampling variation}}.$$

This decomposition shows that the residual stream should not be interpreted as a guaranteed semantically pure private component. It may contain modality-specific information not explained by text, approximation error of the learned conditional distribution, and stochastic variation introduced during reverse sampling. Consequently, the dual-stream diffusion module provides a structural shared–private decomposition, while the downstream task-aware representation learning module determines which components are relevant to the prediction task.

F.6 Acoustic and Visual Instantiations

For the textual-visual stream, the clean target is $\mathbf{x}_0 = \mathbf{V}^\ell$ and the conditioning variable is $\mathbf{T}^\ell$. The learned conditional distribution is $p_{\theta_v}\left(\mathbf{V}^\ell \mid \mathbf{T}^\ell\right)$, from which the shared realization $\widetilde{\mathbf{S}}_v^\ell$ is generated. Its residual representation is

$$\widetilde{\mathbf{P}}_v^\ell = \mathbf{V}^\ell - \widetilde{\mathbf{S}}_v^\ell. \tag{35}$$

For the textual-acoustic stream, the corresponding conditional model is $p_{\theta_a}\left(\mathbf{A}^\ell \mid \mathbf{T}^\ell\right)$, with generated shared representation $\widetilde{\mathbf{S}}_a^\ell$ and residual

$$\widetilde{\mathbf{P}}_a^\ell = \mathbf{A}^\ell - \widetilde{\mathbf{S}}_a^\ell. \tag{36}$$

Stacking the position-wise outputs over $\ell = 1, \ldots, L$ yields $\widetilde{\mathbf{S}}_m, \widetilde{\mathbf{P}}_m \in \mathbb{R}^{L \times d}$ for $m \in \{a, v\}$, which are subsequently used by the task-aware representation learning module.

G Shared-Private Input Construction

This appendix specifies how the sequence-level representations are converted into the shared and private reasoning inputs used in Sec. 3.3. Let $\mathbf{H}^t = \mathbf{T} \in \mathbb{R}^{L \times d}$ denote the textual sequence representation, and let $\mathrm{Pool}(\cdot)$ denote masked mean pooling over valid temporal positions.

G.1 Textual Reasoning Inputs

Since text serves as the conditioning anchor of the dual-stream diffusion module, it does not have a diffusion-derived residual stream. We therefore construct its shared reasoning input directly as

$$\mathbf{u}_t = \mathrm{Pool}\left(\mathbf{H}^t\right). \tag{37}$$

To obtain a complementary textual direction, let $\mathbf{h}_{\text{cls}} \in \mathbb{R}^d$ denote the global textual feature. We remove from $\mathbf{u}_t$ the component aligned with $\mathbf{h}_{\text{cls}}$:

$$\mathbf{v}_t = \mathbf{u}_t - \frac{\mathbf{u}_t^\top \mathbf{h}_{\text{cls}}}{\max\left(\|\mathbf{h}_{\text{cls}}\|_2^2, \varepsilon\right)} \mathbf{h}_{\text{cls}}, \tag{38}$$

where $\varepsilon > 0$ is a numerical stabilization constant. This construction should be interpreted as projection-based residualization rather than a strict semantic shared–private decomposition of text. In particular, $\mathbf{v}_t$ provides a complementary textual direction for compact reasoning, while the original textual representation remains available to the final prediction pathway.

G.2 Non-Textual Reasoning Inputs

For each non-textual modality $m \in \{a, v\}$, Sec. 3.2 produces a text-conditioned shared realization $\widetilde{\mathbf{S}}_m \in \mathbb{R}^{L\times d}$ and the corresponding residual stream $\widetilde{\mathbf{P}}_m \in \mathbb{R}^{L\times d}$. The shared reasoning input is obtained by temporal aggregation:

$$\mathbf{u}_m = \mathrm{Pool}\left(\widetilde{\mathbf{S}}_m\right), \tag{39}$$

where $m \in \{a, v\}$. The residual stream may contain useful modality-specific cues together with conditional-estimation discrepancy and sampling variation. We therefore aggregate it under textual context using attention:

$$\mathbf{v}_m = \mathrm{Pool}\left(\mathrm{MHA}_m^{\mathrm{pr}}\left(\widetilde{\mathbf{P}}_m, \mathbf{H}^t, \widetilde{\mathbf{P}}_m\right)\right), \tag{40}$$

where the three arguments denote the query, key, and value inputs, respectively. Temporal alignment ensures that the textual keys and non-textual residual values refer to corresponding positions. Thus, $\mathbf{u}_m, \mathbf{v}_m \in \mathbb{R}^d, m \in \mathcal{M}$, provide the shared and private inputs to the compact reasoning branch. Because acoustic and visual streams have already been conditioned on text, these vectors are modality-associated representations rather than strictly independent unimodal representations.

H Variational Derivation of the Asymmetric Information-Bottleneck Formulation

This appendix provides the variational derivation of the asymmetric information-bottleneck formulation introduced in Sec. 3.3. For modality $m \in \mathcal{M}$, let U_m and V_m denote the random variables associated with the shared and private reasoning inputs $\mathbf{u}_m$ and $\mathbf{v}_m$, respectively. The compact shared latent Z_m^{sh} is generated by a stochastic encoder $q_{\phi_m}(z_m^{\mathrm{sh}} \mid u_m)$, whereas the compact private latent is obtained deterministically as

$$Z_m^{\mathrm{pr}} = f_m^{\mathrm{pr}}(V_m). \tag{41}$$

Further, the joint compact representation is $Z_m = \left[Z_m^{\mathrm{sh}}; Z_m^{\mathrm{pr}}\right]$. Thus, we can obtain that the constrained information-bottleneck problem in Sec. 3.3 is

$$\min_{\Theta_m} I\left(U_m; Z_m^{\mathrm{sh}}\right) \quad \text{s.t.} I\left(Z_m; Y\right) \geq \kappa. \tag{42}$$

Its Lagrangian relaxation, after omitting the parameter-independent constant $\beta_{\mathrm{IB}}\kappa$, is

$$\mathcal{J}_{\mathrm{IB}}^m = I\left(U_m; Z_m^{\mathrm{sh}}\right) - \beta_{\mathrm{IB}} I\left(Z_m; Y\right). \tag{43}$$

This objective is asymmetric because compression acts only on the shared latent Z_m^{sh}, whereas task relevance is imposed on the joint representation $Z_m = [Z_m^{\mathrm{sh}}; Z_m^{\mathrm{pr}}]$. The private latent therefore contributes to predictive relevance without receiving an additional input-compression penalty.

H.1 Variational Upper Bound on Shared Compression

Proposition 1. Let $r(z_m^{\mathrm{sh}})$ be any reference distribution whose support contains that of the aggregated posterior $q_{\phi_m}(z_m^{\mathrm{sh}})$. Then

$$I\left(U_m; Z_m^{\mathrm{sh}}\right) \leq \mathbb{E}_{p(u_m)}\left[D_{\mathrm{KL}}\left(q_{\phi_m}\left(z_m^{\mathrm{sh}} \mid u_m\right) \,\|\, r\left(z_m^{\mathrm{sh}}\right)\right)\right]. \tag{44}$$

Proof. The aggregated posterior is

$$q_{\phi_m}\left(z_m^{\mathrm{sh}}\right) = \int q_{\phi_m}\left(z_m^{\mathrm{sh}} \mid u_m\right) p(u_m)\, du_m. \tag{45}$$

By definition,

$$I\left(U_m; Z_m^{\mathrm{sh}}\right) = \mathbb{E}_{p(u_m)}\left[D_{\mathrm{KL}}\left(q_{\phi_m}(z_m^{\mathrm{sh}} \mid u_m) \,\|\, q_{\phi_m}(z_m^{\mathrm{sh}})\right)\right]. \tag{46}$$

For an arbitrary reference distribution $r(z_m^{\mathrm{sh}})$,

$$\mathbb{E}_{p(u_m)} D_{\mathrm{KL}}\left(q_{\phi_m}(z_m^{\mathrm{sh}} \mid u_m) \,\|\, r(z_m^{\mathrm{sh}})\right) = I\left(U_m; Z_m^{\mathrm{sh}}\right) + D_{\mathrm{KL}}\left(q_{\phi_m}(z_m^{\mathrm{sh}}) \,\|\, r(z_m^{\mathrm{sh}})\right).$$

Since the second term is nonnegative, Eq. (44) follows. □

We choose the isotropic Gaussian reference $r\left(z_m^{\mathrm{sh}}\right) = \mathcal{N}\left(\mathbf{0}, \mathbf{I}\right)$. Hence, the KL regularizer used in the compact shared branch is a variational upper bound on $I(U_m; Z_m^{\mathrm{sh}})$.

H.2 Variational Lower Bound on Task Relevance

Let E_m denote the random variable induced by the deterministic evidence-construction mapping $\mathbf{e}_m = g_{\mathrm{ev}}\left([\bar{\mathbf{z}}_m^{\mathrm{sh}}; \mathbf{z}_m^{\mathrm{pr}}]\right)$, where $\bar{\mathbf{z}}_m^{\mathrm{sh}}$ is a deterministic transformation of $\mathbf{z}_m^{\mathrm{sh}}$ as defined in Sec. 3.4. Hence, E_m is a deterministic function of Z_m.

Proposition 2. The task information retained by the joint compact representation satisfies

$$I(Z_m; Y) \geq I(E_m; Y) \geq H(Y) + \mathbb{E}_{p(y,e_m)}\left[\log q_\psi(y \mid e_m)\right]. \tag{47}$$

Proof. Since E_m is a deterministic function of Z_m, the Markov chain $Y \to Z_m \to E_m$ implies, by the data-processing inequality, $I(Z_m; Y) \geq I(E_m; Y)$. Moreover,

$$I(E_m; Y) = H(Y) - H(Y \mid E_m). \tag{48}$$

For any variational likelihood $q_\psi(y \mid e_m)$, the nonnegativity of conditional KL divergence gives

$$\mathbb{E}_{p(e_m)} D_{\mathrm{KL}}\left(p(y \mid e_m) \,\|\, q_\psi(y \mid e_m)\right) \geq 0. \tag{49}$$

Therefore,

$$-H(Y \mid E_m) \geq \mathbb{E}_{p(y,e_m)}\left[\log q_\psi(y \mid e_m)\right]. \tag{50}$$

Combining the two inequalities proves the result. □

H.3 Variational Surrogate

Combining Propositions 1 and 2 gives

$$\mathcal{J}_{\mathrm{IB}}^m \leq \mathbb{E}_{p(u_m)} D_{\mathrm{KL}}\left(q_{\phi_m}(z_m^{\mathrm{sh}} \mid u_m) \,\|\, r(z_m^{\mathrm{sh}})\right) - \beta_{\mathrm{IB}} \mathbb{E}_{p(y,e_m)}\left[\log q_\psi(y \mid e_m)\right] - \beta_{\mathrm{IB}} H(Y).$$

Define the auxiliary evidence negative log-likelihood as

$$\mathcal{L}_{\mathrm{ev}} = -\frac{1}{M} \sum_{m \in \mathcal{M}} \mathbb{E}\left[\log q_\psi\left(Y \mid E_m\right)\right]. \tag{51}$$

Using the dimension-normalized KL term defined in the main text, averaging over modalities gives

$$\frac{1}{M} \sum_{m \in \mathcal{M}} \mathcal{J}_{\mathrm{IB}}^m \leq b\mathcal{L}_{\mathrm{KL}} + \beta_{\mathrm{IB}} \mathcal{L}_{\mathrm{ev}} - \beta_{\mathrm{IB}} H(Y). \tag{52}$$

Since $H(Y)$ is independent of the trainable parameters, the KL compression term and the evidence likelihood jointly provide a tractable variational surrogate for the asymmetric compression-relevance trade-off. In practice, the positive scaling factors in this variational surrogate are absorbed into the independently specified coefficients λ_{KL} and λ_{ev} in the overall training objective.

H.4 Closed Form of the Gaussian Compression Term

The compact shared latent is parameterized as

$$q_{\phi_m}\left(z_m^{\mathrm{sh}} \mid u_m\right) = \mathcal{N}\left(\boldsymbol{\mu}_m, \mathrm{diag}(\boldsymbol{\sigma}_m^2)\right). \tag{53}$$

Let $\boldsymbol{\ell}_m = \log \boldsymbol{\sigma}_m^2$. Then

$$D_{\mathrm{KL}} = \frac{1}{2}\sum_{j=1}^{b}\left(\mu_{m,j}^2 + \exp(\ell_{m,j}) - 1 - \ell_{m,j}\right). \tag{54}$$

For a minibatch of size B, the dimension- and modality-normalized implementation is

$$\mathcal{L}_{\mathrm{KL}} = \frac{1}{2MBb}\sum_{m=1}^{M}\sum_{i=1}^{B}\sum_{j=1}^{b}\left(\mu_{mij}^2 + \exp(\ell_{mij}) - 1 - \ell_{mij}\right). \tag{55}$$

This variational information-bottleneck argument applies only to the shared latent in the compact reasoning branch. It does not imply that the representations used for final prediction are information-minimal, because the full-width shared and private representations retained from the text-conditioned decomposition stage remain outside the compact bottleneck and available to the final fusion pathway.

I Task-Conditioned Residual Dependence

We provide the residual-kernel construction and algebraic properties of the task-conditioned dependence regularizer introduced in Sec. 3.3. The purpose of the regularizer is not to enforce unconditional independence between the shared and private representations. Instead, it measures their remaining dependence after regularized removal of variation associated with the task target.

I.1 Residual-Kernel Construction

For a minibatch of size B, let $\mathbf{K}_m^{\mathrm{sh}}$, $\mathbf{K}_m^{\mathrm{pr}}$, and $\mathbf{K}_y \in \mathbb{R}^{B\times B}$ denote centered RBF Gram matrices constructed from the shared posterior means $\boldsymbol{\mu}_m$, private compact latents $\mathbf{z}_m^{\mathrm{pr}}$, and task targets, respectively. The task-conditioned residual operator is

$$\mathbf{R}_y = \mathbf{I}_B - \mathbf{K}_y\left(\mathbf{K}_y + \lambda_r \mathbf{I}_B\right)^{-1}. \tag{56}$$

Applying the same residual operator to both branches gives

$$\widetilde{\mathbf{K}}_m^{\mathrm{sh}} = \mathbf{R}_y \mathbf{K}_m^{\mathrm{sh}} \mathbf{R}_y, \tag{57}$$

and

$$\widetilde{\mathbf{K}}_m^{\mathrm{pr}} = \mathbf{R}_y \mathbf{K}_m^{\mathrm{pr}} \mathbf{R}_y. \tag{58}$$

The normalized residual-kernel dependence coefficient is

$$\rho_m = \frac{\left\langle \widetilde{\mathbf{K}}_m^{\mathrm{sh}}, \widetilde{\mathbf{K}}_m^{\mathrm{pr}} \right\rangle_F}{\left\| \widetilde{\mathbf{K}}_m^{\mathrm{sh}} \right\|_{F,\varepsilon} \left\| \widetilde{\mathbf{K}}_m^{\mathrm{pr}} \right\|_{F,\varepsilon}}. \tag{59}$$

The stabilized Frobenius norm is defined as

$$\|\mathbf{A}\|_{F,\varepsilon} = \sqrt{\max\left(\|\mathbf{A}\|_F^2, \varepsilon\right)}, \quad \varepsilon > 0. \tag{60}$$

I.2 Properties of the Residual Dependence Coefficient

Because $\mathbf{K}_y$ is symmetric positive semidefinite, $\mathbf{R}_y$ is symmetric and can be written spectrally as

$$\mathbf{R}_y = \mathbf{Q}\,\mathrm{diag}\left(\frac{\lambda_r}{\lambda_i + \lambda_r}\right)\mathbf{Q}^\top, \tag{61}$$

where $\{\lambda_i\}$ are the eigenvalues of $\mathbf{K}_y$. Thus, for $\lambda_r > 0$, $\mathbf{R}_y$ attenuates target-associated directions rather than removing them exactly. Since $\widetilde{\mathbf{K}}_m^{\mathrm{sh}}$ and $\widetilde{\mathbf{K}}_m^{\mathrm{pr}}$ remain symmetric positive semidefinite, their Frobenius inner product is nonnegative. Together with the Cauchy-Schwarz inequality, $0 \le \rho_m \le 1$ in exact arithmetic, apart from the numerical stabilization introduced by ε.

Accordingly, the implementation uses $\max(0, \rho_m)$ only as a numerical safeguard against small negative values caused by floating-point errors. It does not alter the intended dependence measure under exact arithmetic. Importantly,

$$\mathbf{R}_y^2 \neq \mathbf{R}_y \tag{62}$$

in general when $\lambda_r > 0$. Therefore, $\mathbf{R}_y$ is not an orthogonal projection and does not completely remove target-associated variation. The resulting objective should consequently be interpreted as a penalty on *residual shared-private dependence after regularized target adjustment*, rather than as a guarantee of exact conditional independence.

J Evidence, Disagreement, and Latent-State Interpretation

This appendix provides the task-specific predictive disagreement definitions, analyzes why disagreement alone is insufficient for determining modality contributions, and gives a latent-state interpretation of the reasoning formulation in Sec. 3.4.

J.1 Evidence Construction and Task Likelihood

The compact shared representation is mapped through a shared residual transformation:

$$\bar{\mathbf{z}}_m^{\mathrm{sh}} = \mathrm{LN}\left(\mathbf{z}_m^{\mathrm{sh}} + \alpha_{\mathrm{sh}} g_{\mathrm{sh}}(\mathbf{z}_m^{\mathrm{sh}})\right), \tag{63}$$

where g_{sh} is shared across modalities and $\alpha_{\mathrm{sh}} \in (0, 1)$ controls the residual mapping. The modality-associated evidence is

$$\mathbf{e}_m = g_{\mathrm{ev}}\left([\bar{\mathbf{z}}_m^{\mathrm{sh}}; \mathbf{z}_m^{\mathrm{pr}}]\right). \tag{64}$$

A parameter-shared prediction head defines $q_\psi(y \mid \mathbf{e}_m)$, whose form follows the downstream task. For regression, we use a fixed-variance Gaussian likelihood, whose negative log-likelihood is equivalent to a squared-error objective up to a positive scale and an additive constant. For C-class classification, q_ψ is instantiated as a categorical likelihood. Thus, the auxiliary objective is

$$\mathcal{L}_{\mathrm{ev}} = -\frac{1}{M}\sum_{m \in \mathcal{M}} \mathbb{E}\left[\log q_\psi\left(Y \mid \mathbf{e}_m\right)\right]. \tag{65}$$

The same prediction mapping is applied to every modality. Hence, evidence comparability is induced through common target supervision rather than direct feature-space correspondence. Its variational relation to $I(Z_m; Y)$ follows from Appendix H.

J.2 Task-Specific Predictive Disagreement

The general disagreement operator in Eq. (11) is instantiated in the downstream task space. For sentiment regression, the auxiliary prediction is bounded to the sentiment range as

$$\widehat{y}_m = 3\tanh\left(h_{\mathrm{ev}}(\mathbf{e}_m)\right), \tag{66}$$

and the normalized pairwise disagreement is

$$D_{mn}^{\mathrm{reg}} = \frac{|\widehat{y}_m - \widehat{y}_n|}{6}. \tag{67}$$

Since $\widehat{y}_m \in [-3, 3]$, this quantity lies in $[0, 1]$. For C-class classification, let

$$\mathbf{p}_m = \mathrm{softmax}\left(h_{\mathrm{ev}}(\mathbf{e}_m)\right) \in \Delta_C \tag{68}$$

denote the predictive class-probability vector. The pairwise disagreement is

$$D^{\mathrm{cls}}_{mn} = \frac{1}{C} \left\| \mathbf{p}_m - \mathbf{p}_n \right\|_1 . \tag{69}$$

Thus, both regression and classification instantiate the general disagreement operator in a task-aware space. The resulting disagreement matrix $\mathbf{D}$ is symmetric and satisfies $D_{mm} = 0$. The stop-gradient operation prevents the final task objective from modifying the auxiliary predictions through the disagreement computation. The evidence representations themselves remain differentiable, while $\mathcal{L}_{\mathrm{ev}}$ provides direct task supervision.

J.3 Non-Identifiability from Disagreement Alone

Proposition. Pairwise predictive disagreement does not, in general, uniquely identify the absolute task interpretation or modality correctness.

Consider the regression disagreement defined in Eq. (67). For any common translation c such that all translated predictions remain within the valid output range, define $\hat{y}'_m = \hat{y}_m + c$. Then,

$$D'_{mn} = \frac{|(\hat{y}_m + c) - (\hat{y}_n + c)|}{6} = D_{mn}. \tag{70}$$

Proof. The common translation cancels from every pairwise difference. □

Consequently,

$$\mathbf{D}(\hat{\mathbf{y}}) = \mathbf{D}(\hat{\mathbf{y}} + c\mathbf{1}), \tag{71}$$

although the absolute locations of the modality predictions in the task space change. Hence, the same relational disagreement structure can correspond to distinct absolute task interpretations. This does not imply that disagreement is uninformative. Rather, it shows that disagreement is relational: without additional task evidence or sample context, it cannot determine which modality prediction is closer to the target. This motivates the latent-state formulation in Sec. 3.4.

J.4 Latent-State Interpretation

Let $\mathcal{E} = \{\mathbf{e}_m\}_{m \in \mathcal{M}}$ denote the modality-associated task evidence and let $\mathbf{D} = [D_{mn}]_{m,n \in \mathcal{M}}$ denote the corresponding predictive disagreement structure. The mapped shared latents are first aggregated into a learned shared context:

$$\gamma_m = \frac{\exp\left(r_{\mathrm{sh}}(\bar{\mathbf{z}}^{\mathrm{sh}}_m)\right)}{\sum_{n \in \mathcal{M}} \exp(r_{\mathrm{sh}}(\bar{\mathbf{z}}^{\mathrm{sh}}_n))}, \tag{72}$$

and

$$\mathbf{c}_{\mathrm{sh}} = \sum_{m \in \mathcal{M}} \gamma_m \bar{\mathbf{z}}^{\mathrm{sh}}_m. \tag{73}$$

Thus, the initial latent task state is

$$\mathbf{h}^{(0)} = f_{\mathrm{init}}(\mathbf{c}_{\mathrm{sh}}). \tag{74}$$

For modality m, let $\mathbf{d}_m = [D_{mn}]_{n \in \mathcal{M}}$ denote its disagreement profile. A preliminary state-conditioned routing distribution is obtained as

$$\widetilde{w}_m = \mathrm{softmax}_m \left(s_\vartheta(\mathbf{h}^{(0)}, \mathbf{e}_m, \mathbf{d}_m) \right), \tag{75}$$

which yields the routed evidence

$$\mathbf{e}_{\mathrm{route}} = \sum_{m \in \mathcal{M}} \widetilde{w}_m \mathbf{e}_m. \tag{76}$$

Together with an encoded disagreement context $\mathbf{c}_D = f_D(\mathbf{D})$, the routed evidence produces a single residual state update:

$$\mathbf{h}^{(1)} = \mathrm{LN} \left(\mathbf{h}^{(0)} + \eta f_{\mathrm{upd}}([\mathbf{e}_{\mathrm{route}}; \mathbf{c}_D]) \right), \tag{77}$$

where $\eta \in (0, 1)$ and further the same routing scorer assigns the state-conditioned contribution score

$$a_m = s_\vartheta(\mathbf{h}^{(1)}, \mathbf{e}_m, \mathbf{d}_m), \tag{78}$$

followed by

$$w_m = \frac{\exp(a_m)}{\sum_{n\in\mathcal{M}} \exp(a_n)}. \tag{79}$$

The preliminary coefficients $\widetilde{w}_m$ are used only to construct the evidence context for state refinement, whereas w_m denotes the final modality contribution used for full-width feature fusion.

J.5 Entropy-Regularized Interpretation of Modality Contributions

Let a_m denote the state-conditioned contribution score in Eq. (13). The softmax normalization

$$w_m = \frac{\exp(a_m)}{\sum_{n\in\mathcal{M}} \exp(a_n)} \tag{80}$$

is equivalently characterized by

$$\mathbf{w} = \arg\max_{\mathbf{q}\in\Delta_M} \left\{ \sum_{m\in\mathcal{M}} q_m a_m + \mathcal{H}(\mathbf{q}) \right\}, \tag{81}$$

where

$$\Delta_M = \left\{ \mathbf{q} \in \mathbb{R}_+^M : \sum_{m=1}^{M} q_m = 1 \right\} \tag{82}$$

and $\mathcal{H}(\mathbf{q}) = -\sum_m q_m \log q_m$.

Thus, the normalized modality contributions can be interpreted as the solution of an entropy-regularized score-maximization problem. This identity characterizes the softmax parameterization only; it does not imply that the learned scores or weights recover ground-truth modality reliability.

K State-Guided Fusion and Learning

This appendix specifies how the modality contributions inferred by the compact reasoning branch are converted into feature gates and applied to the full-width shared and private representations retained from the text-conditioned decomposition stage for final fusion and prediction.

K.1 Feature Gates and Scale Preservation

The normalized modality contributions are converted into feature gates, i.e., $g_m = M w_m$ and $\sum_{m\in\mathcal{M}} g_m = M$. Hence, uniform contributions $w_m = 1/M$ yield $g_m = 1$, so the weighting mechanism preserves the scale of the ungated fusion pathway.

K.2 Gated Shared-Interaction Fusion

Let $\mathbf{r}_t$, $\mathbf{r}_a^{\mathrm{sh}}$, and $\mathbf{r}_v^{\mathrm{sh}}$ denote the pooled full-width textual, acoustic-shared, and visual-shared representations retained from the original feature pathway. After modality-specific projections φ_m, we construct constant-augmented gated factors $\widetilde{\mathbf{r}}_m = [1; g_m \varphi_m(\mathbf{r}_m)]$. Then, the outer product defines the multimodal interaction tensor

$$\mathcal{T} = \widetilde{\mathbf{r}}_t \otimes \widetilde{\mathbf{r}}_a \otimes \widetilde{\mathbf{r}}_v. \tag{83}$$

The shared-interaction representation is

$$\mathbf{f}_{\mathrm{sh}} = f_{\mathrm{ten}}\left(\mathrm{vec}(\mathcal{T})\right), \tag{84}$$

where f_{ten} is the learned tensor-fusion mapping. The constant augmentation retains interaction terms of different orders. In particular, a term involving a modality subset $\mathcal{A} \subseteq \mathcal{M}$ is scaled by $\prod_{m\in\mathcal{A}} g_m$. Thus, the state-conditioned contributions regulate not only individual modality features but also their cross-modal interactions.

K.3 Private-Information Injection

Rather than using the compact bottleneck latents, the private fusion pathway operates on the global textual feature $\mathbf{h}_{\text{cls}}$ and the full-width private acoustic and visual representations $\mathbf{v}_a$ and $\mathbf{v}_v$ retained before compact projection.

$$\mathbf{f}_{\text{pr}} = g_t \mathbf{h}_{\text{cls}} + g_a \mathbf{v}_a + g_v \mathbf{v}_v. \tag{85}$$

Hence, the compact reasoning branch determines *how* each modality contributes, whereas the full-width shared and private representations retained from the decomposition stage provide *what* is fused for final prediction.

K.4 Stabilized Projection-Based Residualization

To reduce overlap between the shared-interaction and private fusion representations, define

$$\Pi^{\varepsilon}_{\mathbf{f}_{\text{pr}}}(\mathbf{f}_{\text{sh}}) = \frac{\mathbf{f}_{\text{sh}}^{\top}\mathbf{f}_{\text{pr}}}{\max\left(\|\mathbf{f}_{\text{pr}}\|_2^2, \varepsilon\right)}\mathbf{f}_{\text{pr}}, \tag{86}$$

where $\varepsilon > 0$. The final fused representation is

$$\mathbf{f} = \mathbf{f}_{\text{pr}} + \mathbf{f}_{\text{sh}} - \Pi^{\varepsilon}_{\mathbf{f}_{\text{pr}}}(\mathbf{f}_{\text{sh}}). \tag{87}$$

When $\|\mathbf{f}_{\text{pr}}\|_2^2 > \varepsilon$, the operator reduces to the standard projection of $\mathbf{f}_{\text{sh}}$ onto the direction $\mathbf{f}_{\text{pr}}$. When stabilization is active, the operation should instead be interpreted as projection-based residualization rather than an exact orthogonal projection.

K.5 Task Prediction and Learning Objective

The final task prediction is $\widehat{Y} = h_{\text{task}}(\mathbf{f})$. Let $\mathbf{X} = (\mathbf{T}, \mathbf{A}, \mathbf{V})$ denote the multimodal input. The primary supervised objective is

$$\mathcal{L}_{\text{task}} = \mathbb{E}_{(\mathbf{X},Y)}\left[\ell_{\text{task}}(\widehat{Y}, Y)\right], \tag{88}$$

where ℓ_{task} is instantiated according to the downstream task. The overall InfoMind objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{diff}}\mathcal{L}_{\text{diff}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{red}}\mathcal{L}_{\text{red}} + \lambda_{\text{ev}}\mathcal{L}_{\text{ev}}. \tag{89}$$

The terms respectively optimize the final supervised task, text-conditioned diffusion decomposition, shared-latent compression, task-conditioned residual dependence, and task-level evidence supervision. The main task objective also trains the latent-state reasoning mechanism through the state-conditioned fusion gates.

During training, stochastic shared bottleneck samples are used in the compact reasoning pathway. At inference, they are replaced by the corresponding posterior means for deterministic prediction.