BTW: A Non-Parametric Variance Stabilization Framework for Multimodal Model Integration

Jun Hou ♠, **Le Wang** ♢, **Xuan Wang** ♠
♠ Department of Computer Science, Virginia Tech, Blacksburg, VA, USA
♢ Department of Agricultural and Applied Economics, Virginia Tech, Blacksburg, VA, USA
{junh, lewangecon, xuanw}@vt.edu

Abstract

Mixture-of-Experts (MoE) models have become increasingly powerful in multimodal learning by enabling modular specialization across modalities. However, their effectiveness remains unclear when additional modalities introduce more noise than complementary information. Existing approaches, like the Partial Information Decomposition, struggle to scale beyond two modalities and lack instance-level control. We propose **B**eyond **T**wo-modality **W**eighting (BTW) [1], a bi-level, non-parametric weighting framework that combines instance-level Kullback-Leibler (KL) divergence and modality-level mutual information (MI) to dynamically adjust modality importance during training. Our method requires no extra parameters and supports any number of modalities. Specifically, BTW computes per-example KL weights by measuring divergence between each unimodal and the current multimodal prediction, and modality-wide MI weights by estimating global alignment between unimodal and multimodal outputs. Extensive experiments on sentiment regression and clinical classification demonstrate that our method significantly improves regression performance and multiclass classification accuracy.

1 Introduction

Multimodal learning has advanced rapidly in vision-language reasoning (Lin et al., 2024a), emotion recognition (Zadeh et al., 2018), and clinical decision support (Soenksen et al., 2022; Hou and Wang, 2025). Sparsely-gated Mixture-of-Experts (MoE) models have emerged as powerful and efficient solutions for scaling multimodal architectures through modular specialization across expert subnetworks (Shazeer et al., 2017; Fedus et al., 2022). Designed for modality-specific and cross-modal patterns, MoE models have achieved success in multimodal fusion, including vision-language grounding, representation learning, and alignment (Feng et al., 2022; Mustafa et al., 2022). Recent MoE frameworks flexibly integrate three or more modalities, achieving strong results across domains (Han et al., 2024; Yun et al., 2024; Li et al., 2025).

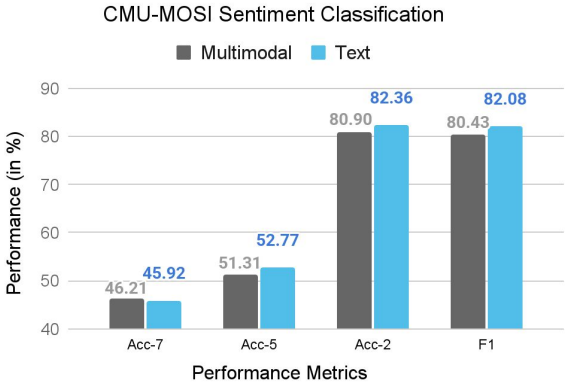


Figure 1: Illustration of our motivation. The CMU-MOSI dataset text modality stand alone could performs better than the multimodal in 5-class classification (Acc-5), binary classification (Acc-2) and Weighted-F1 score.

Despite these advances, it is unclear if integrating multiple modalities adds more noise than useful information, or if current MoE designs capture cross-modal interactions effectively. As shown in Figure 1, the text-only model outperforms the full multimodal system in 5-class and binary sentiment classification (Acc-5, Acc-2), as well as Weighted-F1, on the CMU-MOSI dataset (Zadeh et al., 2016). This result suggests that merely aggregating modalities can hurt performance when one view is substantially more informative, as also seen in healthcare (Hager et al., 2024), highlighting the need for a mechanism that both stabilizes variance and selectively amplifies complementary information to achieve a more effective multimodal integration. Although multiple approaches for modeling modality interactions exist (Table 1), each has no-

[1] https://github.com/JuneHou/Multimodal-Infomax-moe.git

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15089–15103
November 4-9, 2025

Table 1: Comparison of modality interaction methods by scalability, parametric nature, task-specificity, and bi-level design. BTW uniquely supports all four, enabling robust and efficient modeling.

	Scalable	Non-parametric	Task-agnostic	Bi-level
BTW	✔	✔	✔	✔
MI	✔	✔	✔	✘
PID	✘	✔	✘	✘
Game-theory	✔	✘	✔	✘
Attention	✔	✘	✔	✘

table limitations.. Mutual information (MI)-based methods (Shannon, 1948; Han et al., 2021; He et al., 2024) select informative modalities, but lack fine-grained variance control. Partial Information Decomposition (PID) (Williams and Beer, 2010; Liang et al., 2023) provides a principled way to disentangle information among modalities, but is hard to scale up and to generalize to regression tasks. Game-theoretic frameworks (Kontras et al., 2024) and attention mechanisms (Zhang et al., 2023) learn dynamic modality weights but are computationally costly.

To address these limitations, we propose BTW, a novel bi-level weighting mechanism that stabilizes variance across modalities and manages noise from added modalities. At the instance-level, BTW employs Kullback-Leibler (KL) divergence (Kullback, 1997) to measure how much multimodal predictions capture the distributional information from each individual modality. At the modality-level, it leverages MI to quantify global reliability and contribution of each modality's prediction across the dataset. By combining KL and MI, BTW dynamically balances instance-level contributions with global modality reliability, reducing variance while preserving complementary information. Based on experimental results, our framework improves model stability and performance across both continuous (regression) and categorical (classification) tasks, as demonstrated by performance gains in emotion recognition tasks ((Zadeh et al., 2016), (Bagher Zadeh et al., 2018)) and clinical length-of-stay prediction tasks (Johnson et al., 2023).

2 Related Work

2.1 Multimodal Fusion with MoE

Recent multimodal MoE works enhanced fine-grained modality understanding and cross-modal interaction through the use of modality-aware experts (Lin et al., 2024b), local-to-global expert hierarchies (Cao et al., 2023), and text-guided expert activation (Zhao et al., 2024). Other methods are designed so that different modalities selectively guide expert usage, such as the sparsely activated architecture in SkillNet (Dai et al., 2022) and the gating-based dynamic fusion in DynMM (Xue and Marculescu, 2023). These methods aim to dynamically match modality complexity with appropriate expert pathways. To further scale MoE architectures beyond two modalities and for large-scale tasks, recent approaches integrate modality-specific encoders and expert parallelism into unified models. For instance, FuseMoE (Han et al., 2024) introduces *per-modality* and *disjoint* routers to handle heterogeneous modality integration. FlexMoE (Yun et al., 2024) supports scalable integration of any subset of modalities through a unified, flexible expert routing mechanism. Uni-MoE (Li et al., 2025) decouples data and model parallelism across modality-specific experts. IMP (Akbari et al., 2023) leverages alternating gradient descent to integrate multimodal perception efficiently.

2.2 Modality Interaction Modeling

Before modality interactions can be effectively modeled, the missingness handling (Lin and Hu, 2023), contrastive learning (Poklukar et al., 2022) and data augmentation (Lin and Hu, 2024) have improved robustness in multimodal representations. Non-statistical approaches interpret or balance modality contributions using attention-based fusion (Tsai et al., 2019; Zhang et al., 2023) and gradient-based visualization (Chen et al., 2023). Statistical interaction modeling is favored for its model-agnostic, non-parametric nature and is applied at either global or instance levels.

At the global level, MI has been used to model modality interactions (Han et al., 2021; He et al., 2024), and the information bottleneck has been used to reduce noise (Wu et al., 2023). But these lack instance-level or directional specificity, limiting use for heterogeneous data. PID (Williams and Beer, 2010; Liang et al., 2023) decomposes information into unique and shared contributions from each modality but is typically limited to two modalities by computational cost. Game-theoretic frameworks (Kontras et al., 2024) further leverage mutual information decomposition to balance modality influence across the dataset. Although these frameworks are generalizable to high-dimensional settings, the assumption of modality competition

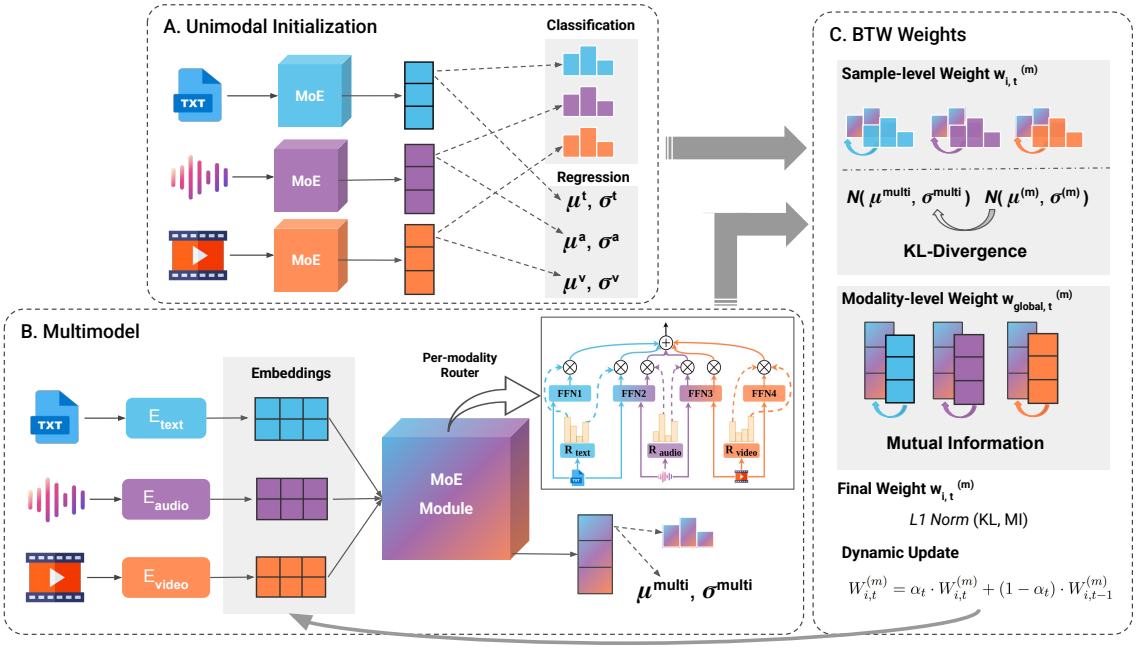


Figure 2: The overall architecture of the proposed BTW weighting framework. (A) Unimodal Initialization: Each modality is processed separately through the shared MoE backbone to produce unimodal predictions $\hat{y}_i^{(m)}$, yielding a Gaussian distribution $(\mu_i^{(m)}, \sigma_i^{(m)})$ for regression and categorical probabilities for classification. (B) Multimodal: All embeddings of each modality are fused by a MoE module with the per-modality routers, producing a prediction $(\mu_i^{multi}, \sigma_i^{multi})$ for regression or categorical probabilities for classification. (C) BTW Weights: The instance-level and modality-level weights are computed based on the predictions from (A) and (B). The final L1-normalized weights $W_i^{(m)}$ are dynamically smoothed across epochs to rescale modality embeddings during training.

tends to overlook the need to resolve conflicts between modalities. At the instance level, DIME (Lyu et al., 2022) attributes model predictions to each modality for individual samples, though its computational complexity can be a limitation. Recent work uses information bottleneck for instance-level modality contribution (Fang et al., 2024), filtering noise rather than modeling interactions.

3 Method

Our proposed BTW framework, as shown in Figure 2, specifically addresses the need for a generalizable solution capable of scaling to arbitrary modalities by systematically analyzing modality interactions while simultaneously stabilizing variance. The framework is built on top of existing MoE models (Han et al., 2024) and involves three steps: (1) obtaining unimodal predictions, (2) computing bi-level weights, and (3) dynamically applying these weights to modality embeddings.

3.1 Unimodal Predictions Initialization

To begin, we aim to extract the maximum amount of information from the input embeddings of each individual modality $X_i^{(m)}$ in the complete data X_i, where $m \in \{1, \ldots, M\}$ and M is the total number of modalities. Consider a multimodal classifier f capable of handling an arbitrary number of modalities. To establish a baseline measure of information provided by each individual modality, we first generate unimodal predictions $\hat{y}_i^{(m)}$ by training f using only one modality at a time:

$$\hat{y}_i^{(0)} = f^{(0)}(X_i^{(0)}) \quad (1)$$

$$\hat{y}_i^{(1)} = f^{(1)}(X_i^{(1)}) \quad (2)$$

$$\vdots$$

$$\hat{y}_i^{(M)} = f^{(M)}(X_i^{(M)}). \quad (3)$$

Next, we train the same model using all available modalities jointly with existing MoE models, without applying any modality-specific weighting:

$$\hat{y}_i^{\text{multi}} = f^{\text{multi}}(X_i). \quad (4)$$

After training the unimodal and multimodal, we collect the model prediction $\hat{y}^{(m)}$ and $\hat{y}^{\text{multi}}$. These unimodal and multimodal predictions serve as essential reference points for quantifying the individual contributions and variability of each modality in subsequent training phases.

3.2 BTW: Instance-Level Weights

Instance-level weights are computed based on KL divergence between the probability distributions of unimodal and multimodal predictions.

Classification tasks can directly output the probability, which for each modality and instance. For modality m, the instance-level weight is calculated as the KL-divergence between the unimodal and multimodal prediction distributions:

$$w_i^{(m)} = D_{KL}(P(\hat{y}_i^{(m)}|X_i^{(m)}) \parallel P(\hat{y}_i^{\text{multi}}|X_i^{\text{multi}})). \tag{5}$$

Regression tasks have continuous output from each modalities prediction. First, unimodal predictions, we model each modality's prediction as a Gaussian distribution with mean $\mu_i^{(m)}$ and variance $\sigma_i^{(m)}$. The network output is interpreted as the conditional mean, $\mu_i^{(m)} = \mathrm{E}[\hat{y}_i^{(m)} \,|\, X_i^{(m)}]$. In addition, we use the squared error $\sigma_i^{(m)} = \left(Y_i - \mu_i^{(m)}\right)^2$ as an estimator of the conditional variance, where Y_i is the ground truth. Second, for multimodal predictions, μ_i^{multi} and σ_i^{multi} are estimated using the same procedure. The unimodal estimations and the multimodal estimations are used in the closed-form KL divergence between the unimodal Gaussian $\mathcal{N}(\mu_i^{(m)}, \sigma_i^{(m)})$ and multimodal Gaussian $\mathcal{N}(\mu_i^{\text{multi}}, \sigma_i^{\text{multi}})$, respectively (see Appendix C for the full derivation). The instance-level weight is calculated as the KL-divergence between the above two Gaussian distributions:

$$w_i^{(m)} = D_{\text{KL}}\big(\mathcal{N}(\mu_i^{(m)}, \sigma_i^{(m)}) \,\|\, \mathcal{N}(\mu_i^{\text{multi}}, \sigma_i^{\text{multi}})\big). \tag{6}$$

These instance-level KL divergences are normalized across modalities for each instance to ensure comparability. A larger KL divergence indicates stronger disagreements between the unimodal and multimodal predictions, indicating more unique information can be learned from this modality, and thus a higher weight in the final integration,

3.3 BTW: Modality-Level Weights

Modality-level weights are designed to quantify global modality reliability and informativeness. We calculate MI between unimodal and multimodal predictions across the entire dataset. Formally, let $\hat{y}^{(m)} = \{\hat{y}_i^{(m)}\}_{i=1}^N$ and $\hat{y}^{\text{multi}} = \{\hat{y}_i^{\text{multi}}\}_{i=1}^N$ denote the predicted outputs from the unimodal model for modality m and the multimodal model, respectively. We define their mutual information as follows (see Appendix E for details):

$$\text{MI}(\hat{y}^{(m)}, \hat{y}^{\text{multi}}) = \sum_{\hat{y}^{(m)}, \hat{y}^{\text{multi}} \in \mathcal{Y}} P(\hat{y}^{(m)}, \hat{y}^{\text{multi}}) \log \frac{P(\hat{y}^{(m)}, \hat{y}^{\text{multi}})}{P(\hat{y}^{(m)}) P(\hat{y}^{\text{multi}})}. \tag{7}$$

In classification tasks, both $\hat{y}^{(m)}$ and $\hat{y}^{\text{multi}}$ are discrete class predictions, each taking values in the set of all possible classes $\mathcal{Y}$, and mutual information is computed between each modality's predicted values and the multimodal prediction across the dataset[2]. For regression tasks, where the predictions are continuous-valued scores, the summation in Eq. 7 is replaced with double integration over $\hat{y}^{(m)}$ and $\hat{y}^{\text{multi}}$. In the implementation mutual information is estimated using non-parametric entropy estimators based on k-nearest neighbor statistics[3]. In both cases, the resulting MI score captures how much predictive information the unimodal modality shares with the multimodal model.

3.4 Dynamically Adapted Bi-level Weights

The modality-level MI weights are used to rescale the instance-level KL divergence weights, amplifying contributions from globally informative modalities while attenuating the influence of less reliable ones. We evaluate two versions of our BTW framework, **BTW-local (KL)** defined as using instance-level weight only, and **BTW** defined as using bi-level weights. At the training epoch t, two versions of final weights are computes as:

$$\text{BTW-local (KL):} \quad W_{i,t}^{(m)} = \frac{w_{i,t}^{(m)}}{\sum_{j=1}^M w_{i,t}^{(j)}}. \tag{8}$$

$$\text{BTW:} \quad W_{i,t}^{(m)} = \frac{w_{i,t}^{(m)} \cdot \text{MI}(\hat{y}^{(m)}, \hat{y}^{\text{multi}})}{\sum_{j=1}^M w_{i,t}^{(j)} \cdot \text{MI}(\hat{y}^{(j)}, \hat{y}^{\text{multi}})}. \tag{9}$$

We dynamically update the computed bi-level weights throughout training epochs based on model performance improvements, measured by the F1 score for classification tasks or mean absolute error (MAE) for regression tasks. A smoothing factor (α_t) with time step t is adaptively adjusted, incremented if performance improves and otherwise

[2] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html

[3] https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.mutual_info_regression.html

decremented, avoiding rapid fluctuations. Specifically, the updated weights $W_{i,t}^{(m)}$ for modality at instance for the epoch are computed as:

$$W_{i,t}^{(m)} = \alpha_t \cdot W_{i,t}^{(m)} + (1 - \alpha_t) \cdot W_{i,t-1}^{(m)}. \quad (10)$$

These stabilized weights are then multiplicatively applied to modality embeddings prior to each subsequent training epoch, emphasizing reliable and informative modalities while reducing variance across individual instances and modalities, thus enhancing robustness and generalization of the multimodal integration.

4 Experiment

We conduct experiments to validate BTW in stabilizing variance and evaluating robustness across modalities, tasks, and diverse domains. Details on computation, time cost, and hyper-parameters are in Appendix F.

4.1 Datasets

We conduct experiments on three publicly available benchmark datasets with more than two modalities:

CMU-MOSI (Zadeh et al., 2016) contains 2,199 clips collected from YouTube opinion videos labeled with sentiment in the range of -3 (negative) and +3 (positive).

CMU-MOSEI (Zadeh et al., 2018): 23,500 clips labeled from sentiment intensity. Both datasets are benchmarks for multimodal sentiment analysis with textual, acoustic and visual modalities.

MIMIC-IV (Johnson et al., 2023) is a large-scale clinical dataset containing rich multimodal patient data, including irregularly sampled time series (e.g., vital signs, lab tests), clinical notes, chest X-ray (CXR), and ECG signals. We adopt the dataset curation pipeline from FuseMoE (Han et al., 2024), in which only 25% of samples include CXR modality and 52% include ECG modality, while all samples contain time-series modality. For our experiments, we focus on predicting patient length-of-stay (LOS), re-framed as a 4-class classification task based on clinical grouping criteria proposed by CORe (van Aken et al., 2021). To evaluate the impact of missing modalities on our weighting framework, we construct a subset containing only instances with complete modality availability (statistics summarized in Table 2).

LOS in days	≤3	3–7	7–14	>14	Total
Count	1,465	2,342	923	448	5,178

Table 2: Length-of-stay (LOS) class distribution for the MIMIC-IV dataset used in our experiments. Only the no-missing-modality subset is used for training and evaluation, containing 5,178 complete patient stays.

4.2 Experimental Setup

Baseline MoE Model For all experiments, we adopt the MoE architecture from FuseMoE (Han et al., 2024) as the backbone for multimodal fusion. We select the *per-modality* router, the best-performing strategy in FuseMoE. The per-modality router distributes each modality independently to a shared pool of experts, offering a principled balance between modality-specific specialization and cross-modal integration. Implementation details and hyper-parameters are provided in Appendix F.

Weight Initialization and Adaptation For both tasks, we follow the three-step procedure described in Section 3. After each training epoch, the computed bi-level weights are dynamically updated according to model performance (F1 score for classification, MAE for regression). A smoothing factor (α_t) stabilizes these updates.

Evaluation Metrics Regression results of the MOSI and MOSEI datasets are evaluated using MAE. Additionally, following established sentiment benchmarks (Han et al., 2021), we report Pearson correlation (Corr), seven-class accuracy (Acc-7), five-class accuracy (Acc-5), binary classification accuracy (Acc-2) and F1 score computed for positive/negative and non-negative/negative classification results. Model performance for LOS classification is evaluated using overall Accuracy, Macro F1-score, and Weighted F1-score.

5 Main Results

Sentiment Analysis Table 3 summarizes evaluation results for multimodal sentiment regression across two benchmarks: CMU-MOSI and CMU-MOSEI. We compare our proposed bi-level weighting framework (BTW-local (KL) and BTW) against recent state-of-the-art multimodal fusion methods, including MulT (Tsai et al., 2019), MMIM (Han et al., 2021), and standard MoE (Han et al., 2024). Specifically, on CMU-MOSI, the BTW-local (KL) approach achieves the lowest MAE of 0.714 (-2% improvement over MMIM and MoE) and the

Table 3: Main results on CMU-MOSI and CMU-MOSEI for multimodal sentiment regression. BTW-global (KL), BTW-global (MI), BTW-local (KL) and BTW denote variants of our proposed bi-level weighting framework. Bold numbers indicate the best performance and underlined numbers indicate the second-best. All results are averaged over three runs with different random seeds. MMIM (Han et al., 2021) and MulT (Tsai et al., 2019) results are reproduced from open-source code with the hyperparameters specified.

	Method	MAE↓	Corr	Acc-7	Acc-5	Acc-2	Weighted-F1
MOSI	MulT	0.989±0.04	0.646±0.03	33.33±1.24	36.60±1.77	77.05±1.21 / 78.46±1.42	77.01±1.12 / 78.53±1.36
	MMIM	0.738±0.01	0.778±0.01	44.12±1.31	50.29±0.53	82.51±0.38 / 84.45±1.07	82.34±0.41 / 84.37±0.93
	MoE	0.735±0.00	0.770±0.01	45.87±0.59	52.28±0.89	81.78±1.05 / 83.99±1.07	81.56±1.18 / 83.88±1.02
	BTW-global (KL)	0.746±0.011	0.774±0.001	44.56±1.80	51.65±1.18	82.17±0.37 / 84.00±0.27	82.07±0.42 / 83.96±0.30
	BTW-global (MI)	0.726±0.012	0.776±0.003	44.51±0.89	51.80±1.10	82.34±1.06 / **84.56**±0.31	82.49±0.87 / **84.66**±0.23
	BTW-local (KL)	**0.714**±0.01	**0.786**±0.01	46.40±3.23	53.26±3.31	82.46±0.97 / 84.55±0.92	82.33±0.96 / 84.50±0.88
	BTW	0.716±0.01	0.781±0.01	**47.52**±0.77	**54.28**±1.43	**82.75**±1.17 / 84.35±0.84	**82.68**±1.26 / 84.34±0.91
MOSEI	MulT	0.613±0.01	0.669±0.02	49.55±0.49	50.93±0.64	78.22±0.37 / 80.36±1.40	78.55±0.21 / 80.42±0.89
	MMIM	0.578±0.01	0.728±0.01	51.03±0.42	52.39±0.54	81.61±2.37 / 83.31±0.39	81.23±2.67 / 82.98±0.71
	MoE	0.570±0.01	0.723±0.01	52.17±0.61	53.73±0.64	80.02±3.84 / 83.41±1.12	80.53±3.34 / **83.29**±1.11
	BTW-global (KL)	0.572±0.011	0.725±0.010	52.24±0.44	53.55±0.49	80.90±3.83 / 83.32±0.69	81.14±3.19 / 83.10±0.61
	BTW-global (MI)	**0.566**±0.006	**0.729**±0.002	52.34±0.59	53.95±0.63	76.09±3.40 / 81.78±1.88	75.66±4.16 / 81.68±2.08
	BTW-local (KL)	**0.566**±0.01	0.727±0.01	52.32±0.76	53.85±0.73	**83.02**±1.10 / **83.60**±1.87	**82.81**±0.98 / 83.07±2.29
	BTW	0.573±0.01	0.722±0.00	**52.62**±0.74	**54.15**±0.87	81.97±2.36 / 81.92±0.96	81.50±1.98 / 81.22±1.36

Table 4: Main results on the MIMIC-IV dataset for length-of-stay classification. BTW-local (KL) and BTW denote variants of our bi-level weighting framework. Bold values indicate the best results across all methods. All metrics are averaged over three random seeds.

Method	Accuracy	Macro-F1	Weighted-F1
MulT	43.33±2.08	33.33±3.21	41.33±0.58
HAIM	**46.00**±0.00	33.00±0.00	42.00±0.00
FuseMoE	41.33±1.53	**37.67**±2.89	40.33±2.31
BTW-local (KL)	43.67±2.31	37.00±1.00	43.00±1.73
BTW	45.67±0.58	**37.67**±0.58	**45.00**±0.00

highest Pearson correlation of 0.786, outperforming all baseline models. Incorporating global MI weighting (BTW) yields competitive performance following BTW-local (KL) in regression metrics (MAE=0.716, Corr=0.781). Notably, BTW also achieves the best 7-class accuracy (47.52%), 5-class accuracy (54.28%), and the highest binary accuracy when including the zero-threshold (82.75%). Overall, the BTW-local (KL) weights consistently improve the regression metrics, while BTW presents consistent out-performance in classification tasks.

For the larger CMU-MOSEI dataset, BTW-local (KL) consistently shows strong regression results, achieving an MAE of 0.566, best binary classification accuracy (83.02%/83.60%) and weighted-F1 (82.81% with zero-threshold included). Meanwhile, BTW is consistently leading in multi-class classification, reporting the highest accuracies in 7-class (52.62%) and 5-class (54.15%) settings. Although, BTW demonstrates reduced binary classification accuracy and F1 scores, its outstanding performance in multi-class scenarios highlights the complementary value of incorporating modality-level mutual information.

Length-of-Stay Prediction We evaluate our proposed bi-level weighting framework on the MIMIC-IV dataset for the clinically relevant task of four-class length-of-stay (LOS) classification. For these experiments, we utilize only stays with complete modality data, thereby minimizing modality incompleteness as a confounding factor and focusing exclusively on evaluating the core modality interaction capacity of our weighting methods. Missing modality and 3-modality scenarios are explored separately in Appendix A and Appendix B, respectively. Table 4 summarizes classification results comparing our proposed BTW-local (KL) and BTW weighting variants against three strong baselines: the tree-based machine learning model HAIM (Soenksen et al., 2022), the transformer-based fusion model MulT (Tsai et al., 2019), and the recent FuseMoE (Han et al., 2024). Our BTW weighting method achieves the highest Macro-F1 and significantly outperforms all other methods in Weighted-F1 (+5% over FuseMoE, +4% over MulT, and +3% over HAIM). Additionally, BTW yields the second-highest Accuracy (45.67%, close to the top-performing HAIM model at 46.00%). The instance-level BTW-local (KL) attains the second-

best Macro-F1 (37%) and Weighted-F1 (43%), exceeding all baselines, and achieves competitive Accuracy (43.67%). These results highlight that our bi-level weighting approach substantially enhances multimodal classification performance, effectively balancing class-specific performance and overall accuracy compared to state-of-the-art baselines.

6 Ablation Study

6.1 Weighting Mechanisms

To better understand the individual contributions of each weighting component in our BTW framework, we conducted ablation experiments on the sentiment analysis task. Specifically, for BTW-global (KL), we modify Eq. 8 by replacing the instance-level weight for each modality with its average across the dataset, thus using a constant modality-specific weight for all samples. For BTW-global (MI), we make an analogous change to Eq. 9, using only the global MI value for each modality and omitting instance-level weights. Table 3 summarizes the results of these ablations.

BTW-global (MI) consistently outperforms BTW-global (KL), highlighting the effectiveness of MI in capturing global modality alignment. Compared with BTW-global (KL), BTW-local (KL) demonstrates a clear advantage of instance-level weighting across nearly all metrics on both datasets, underscoring the benefits of fine-grained variance stabilization at the instance-level. Lastly, BTW reveals its complementary strengths by consistently outperforming the BTW-global (MI) in multiclass-classification accuracy (e.g., CMU-MOSEI Acc-7: 52.62% vs. 52.34%), indicating the synergy between local variance management and global modality informativeness.

In summary, KL provides effective instance-level variance control beneficial to regression, while MI contributes significantly to global alignment and multi-class accuracy. The bi-level weights yields balanced, robust multimodal fusion.

6.2 Weight Distribution Analysis

To better understand how our BTW framework stabilizes variance in heterogeneous datasets, we analyze modality-specific weight trajectories across training epochs on MIMIC-IV (Figure 3).

In the BTW-local (KL) case, weights quickly stabilize after the first epoch. Notably, the weights for time-series modality (TS) drop sharply and remain relatively stable thereafter, indicating that the multimodal prediction for TS closely matches what can be achieved with TS alone. This observation aligns with our expectation that TS is inherently easier for the model to capture. However, the high KL divergence weights for text, imaging (CXR), and ECG suggest that these modalities provide distinct, potentially noisy information that requires careful reconciliation during multimodal fusion.

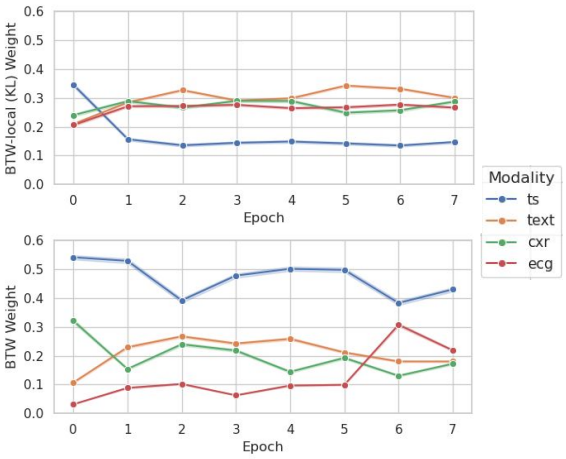


Figure 3: The evolution of modality weights across eight training epochs on MIMIC-IV dataset. The upper panel shows the BTW and the lower panel shows BTW-local (KL) weights. BTW effectively balances the dominant TS modality from overy emphasized the uniqueness.

Incorporating modality-level MI weights reshapes the distribution, consistently emphasizing the globally informative TS while balancing distinct contributions from other modalities, significantly improving accuracy and F1 scores (Table 4). Without weights, the model treats modalities equally, whereas BTW-local (KL) weights overly emphasize the uniqueness of TS. Integrating MI provides a necessary global perspective, balancing the strong predictive power of TS with the unique contributions of other modalities, resulting in a more robust multimodal fusion.

6.3 Encoder Sensitivity

To assess sensitivity to the language encoder, we replace the baseline BERT encoder with DeBERTa while keeping all other components identical, and compare against ITHP and other baselines (Table 5). BTW-local provides the strongest regression metrics, while BTW leads multi-class classification. On MOSI, BTW-local attains MAE=0.691; BTW achieves the best multiclass accuracy and high binary accuracy when including the zero-threshold (82.80%). Similar trends hold on MOSEI, where

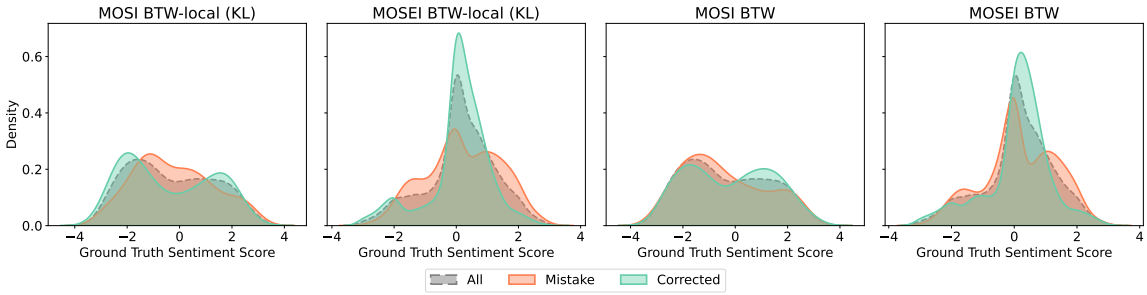


Figure 4: Kernel density plots of ground-truth sentiment scores for the test set, showing the distributions of all samples, predictions corrected by two variations of weighting schemes, and new mistakes introduced, across both MOSI and MOSEI datasets. Both schemes tend to correct errors in high-density regions of the score distribution, with BTW especially concentrating corrections near neutral sentiment.

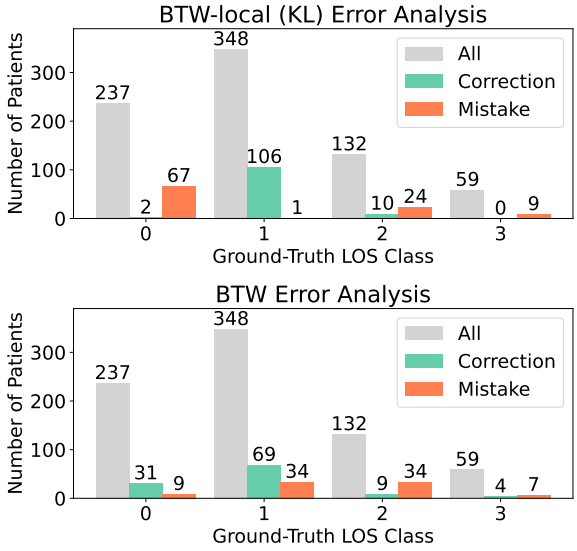


Figure 5: Visualization of test split counts of all instances, corrections, and mistakes over the ground-truth classes for the LOS classification task. BTW distributes corrections more evenly across all classes, notably improving performance in class 0.

BTW-local yields competitive MAE and BTW attains leading Acc-7/Acc-5.

These results show that our framework captures fine-grained sentiment distinctions better than ITHP, which compresses information toward a dominant modality. In contrast, BTW balances all modalities using instance-level KL divergence and modality-level mutual information, preserving diverse signals and improving complex multi-class performance. This demonstrates that our weighting approach yields consistent improvements that are not dependent on the choice of encoder.

6.4 Smoothing Factor (α_t)

We analyze the impact of EMA smoothing factor $\alpha_t \in 0.3, 0.5, 0.7$ on BTW-local and BTW on MOSI/MOSEI. As shown in Table 6, lowering (0.3) or raising (0.7) improves some metrics while reducing others, so we set the default value to $\alpha_t = 0.5$, which consistently offers the best and most stable trade-off across datasets and metrics.

6.5 Case Study: Error Analysis

To further validate how our proposed bi-level weighting framework improves multimodal predictions, we conduct an in-depth case study examining the instances where our methods correct baseline errors or introduce new mistakes. Kernel density estimates of ground-truth scores for all instances, corrections, and new mistakes are plotted for sentiment analysis regression (Figure 4; complementary scatter plots in Appendix D) and LOS classification tasks (Figure 5).

When applied to both tasks, both weighting schemes consistently correct predictions in regions with high data density in the regression task, as higher density implies greater agreement and stronger confidence in corrections via KL divergence. For sentiment analysis, incorporating global MI leads to corrections concentrated around neutral sentiment scores in MOSI and MOSEI. For LOS classification, the results highlight a critical trade-off and the strength of our bi-level approach. While BTW-local (KL) focuses its corrections on the high-density majority class (106 corrections in class 1), it performs poorly on the most challenging minority class (only 2 corrections in class 0). In contrast, the full BTW framework uses global MI to re-balance its focus. It strategically sacrifices some corrections on the majority class to make significant gains in the most difficult classes, most notably increasing corrections in class 0 from 2 to 31. This ability to improve performance on minority classes by resolving uncertainty and ambiguity

Table 5: Performance comparison on CMU-MOSI and CMU-MOSEI using the DeBERTa text encoder. Results are compared against baseline models and the ITHP model. Bold numbers indicate the best performance and underlined numbers indicate the second-best. All results are averaged over three runs with different random seeds. The 'Acc-2' and 'Weighted-F1' columns show 'include-zero / non-zero' results.

	Method	MAE↓	Corr↑	Acc-7↑	Acc-5↑	Acc-2↑	Weighted-F1↑
MOSI	ITHP	0.719±0.004	**0.840**±0.006	43.75±0.75	54.26±1.22	**85.69**±1.02 / **87.33**±1.00	**85.66**±1.05 / **87.34**±1.02
	MMIM	0.713±0.007	0.799±0.005	45.00±0.99	52.72±1.00	82.75±0.17 / 84.93±0.37	82.64±0.23 / 84.77±0.49
	MoE	0.701±0.011	0.793±0.013	44.70±1.09	52.48±1.19	82.46±0.89 / 84.80±0.93	82.30±0.96 / 84.73±0.97
	BTW-local	0.696±0.006	0.801±0.005	45.14±2.09	52.63±1.65	83.53±0.67 / 85.77±1.07	83.38±0.66 / 85.70±1.07
	BTW	**0.691**±0.005	0.790±0.013	**46.55**±1.38	**54.86**±1.35	82.80±0.50 / 85.37±0.16	82.70±0.46 / 85.04±0.62
MOSEI	ITHP	0.599±0.004	**0.782**±0.003	48.30±0.28	52.32±0.32	80.67±0.39 / **85.66**±0.32	81.36±0.34 / **85.75**±0.29
	MMIM	0.566±0.014	0.732±0.020	51.83±0.67	53.11±0.80	80.58±1.86 / 83.34±1.04	80.77±1.35 / 83.03±1.42
	MoE	0.559±0.006	0.742±0.018	52.32±0.46	53.72±0.78	81.16±2.62 / 83.90±1.64	81.39±1.95 / 83.68±2.04
	BTW-local	**0.540**±0.005	0.757±0.005	**53.89**±0.19	**55.63**±0.21	81.75±1.23 / 85.11±0.32	82.10±0.92 / 84.97±0.50
	BTW	0.545±0.007	0.756±0.004	53.35±0.27	55.04±0.16	**82.10**±2.96 / 84.71±0.45	**82.34**±2.47 / 83.48±1.85

Table 6: Ablation study on the smoothing factor (α_t) for BTW-local (KL) and BTW methods on the MOSI and MOSEI datasets. The default value of 0.5 consistently provides the best or most stable performance. All results are averaged over three runs.

	Method	α_t	MAE↓	Corr↑	Acc-7↑	Acc-5↑	Acc-2↑	Weighted-F1↑
MOSI	BTW-local	0.3	0.740±0.016	0.774±0.008	45.48±1.055	52.53±1.752	81.88±0.670/83.94±0.919	81.74±0.688/83.88±0.952
		0.5	**0.714**±0.011	**0.786**±0.005	**46.40**±3.227	**53.26**±3.312	**82.46**±0.970/**84.55**±0.919	**82.33**±0.961/**84.50**±0.883
		0.7	0.758±0.006	0.766±0.003	43.64±1.091	50.19±1.701	81.29±0.734/83.54±0.531	80.94±0.960/82.65±1.473
	BTW	0.3	0.739±0.021	0.771±0.011	45.73±1.097	52.67±1.604	81.54±1.091/83.79±1.424	81.35±0.865/83.54±1.427
		0.5	**0.716**±0.008	**0.781**±0.009	**47.52**±0.773	**54.28**±1.431	**82.75**±1.173/**84.35**±0.836	**82.68**±1.260/**84.34**±0.910
		0.7	0.744±0.013	0.770±0.006	44.95±1.244	52.19±1.197	81.97±0.656/84.60±0.703	81.64±0.606/84.39±0.650
MOSEI	BTW-local	0.3	0.574±0.007	0.721±0.004	52.14±0.440	53.73±0.188	82.83±0.710/82.90±0.613	82.49±0.490/82.26±0.765
		0.5	**0.566**±0.008	**0.727**±0.006	**52.32**±0.757	**53.85**±0.732	**83.02**±1.098/**83.60**±1.869	**82.81**±0.983/**83.07**±2.288
		0.7	0.563±0.003	0.729±0.003	52.69±0.428	54.25±0.359	81.21±4.540/82.64±0.538	82.17±1.989/82.17±0.386
	BTW	0.3	0.574±0.002	0.718±0.005	52.07±0.363	53.69±0.174	81.56±1.712/82.43±0.771	81.41±1.642/81.50±1.525
		0.5	**0.573**±0.007	**0.722**±0.003	**52.62**±0.737	**54.15**±0.867	**81.97**±2.356/**81.92**±0.961	**81.50**±1.976/**81.22**±1.360
		0.7	0.577±0.008	0.716±0.006	52.03±0.397	53.12±0.817	82.41±1.025/82.55±0.882	82.21±0.856/82.04±0.760

in classification boundaries is critical for building a robust and clinically useful model.

While the BTW-local (KL) weights efficiently reduce variance in densely populated sentiment regions, global MI effectively targets the ambiguity regions, underscoring the complementary strengths of the bi-level weights.

7 Conclusion

This paper introduces a bi-level, non-parametric weighting framework that advances multimodal learning beyond two modalities by addressing prediction variance and modality interaction explainability. By integrating instance-level KL divergence with modality-level mutual information, the method adaptively calibrates modality contributions without introducing additional trainable parameters. The framework's value lies not only in quantitative improvements but also in its efficiency as a non-parametric, plug-and-play module that enhances existing architectures. Extensive experiments on diverse benchmarks and both regression and classification tasks demonstrate how the framework stabilizes variance in high-density regions and resolves ambiguity by leveraging globally informative modalities, ultimately facilitating more robust and transparent multimodal models.

Acknowledgement

Our work is sponsored by NSF #2442253, NAIRR Pilot with PSC Neocortex and NCSA Delta, Commonwealth Cyber Initiative, Children's National Hospital, Fralin Biomedical Research Institute (Virginia Tech), Sanghani Center for AI and Data Analytics (Virginia Tech), Virginia Tech Innovation Campus, and generous gifts from Nivida, Cisco, and the Amazon + Virginia Tech Center for Efficient and Robust Machine Learning.

Limitations

While the BTW framework offers improved interpretability and performance across diverse multimodal tasks, there are some limitations that might obstacle further generalization and effectiveness. First, the information theory-based weights rely on the assumption that unimodal predictions provide informative and well-calibrated distributions. In cases where modalities have low quality or completeness, the resulting weights may introduce noise rather than stabilize variance. Future work could explore uncertainty-aware regularization or confidence-based gating to downweight unreliable unimodal predictions.

First, the information theory-based weights rely on the assumption that unimodal predictions provide informative and well-calibrated distributions. In cases where modalities have low quality or completeness, the resulting weights may introduce noise rather than stabilize variance. Future work could explore uncertainty-aware regularization or confidence-based gating to downweight unreliable unimodal predictions.

Second, the BTW framework, in its current form, is demonstrated on an MoE architecture. Its core requirement is the ability to obtain separate predictions from each unimodal path as well as a joint multimodal prediction. Therefore, it is directly applicable to various late-fusion or hybrid-fusion architectures but is not suited for pure early-fusion models where raw features are concatenated at the input layer, preventing the generation of distinct unimodal outputs from the fused representation.

Third, our method inherits the zero-embedding strategy from the FuseMoE backbone (Han et al., 2024) for handling missing modalities. As shown in our ablation, information-theoretic metrics like mutual information fail to provide meaningful signals when modalities are absent, due to the degenerate nature of zero vectors. Risks might arise if modality imputation is inaccurate or missing data are handled improperly, leading to unreliable or misleading model predictions. This suggests that imputing missing modality embeddings with synthetically generated representations could offer a more coherent and informative approximation, preserving the multimodal distributional structure.

Finally, while our method is task-agnostic for regression and classification tasks, supervision through labels is required for the current framework. This supervised assumption limits generalizability to unsupervised or self-supervised settings. Future research could explore proxy objectives such as contrastive similarity or mutual predictability to extend this framework to representation learning.

Ethics Statement

All datasets used in this research are publicly available for research use, under the terms and licenses specified by their creators. MIMIC-IV is released under the PhysioNet Credentialed Health Data Use Agreement [4]. No proprietary data or restricted-access resources were used. We do not display raw excerpts from any dataset in this paper. We do not attempt to identify or deanonymize users in the data in any way during our research.

References

Hassan Akbari, Dan Kondratyuk, Yin Cui, Rachel Hornung, Huisheng Wang, and Hartwig Adam. 2023. Alternating gradient descent and mixture-of-experts for integrated multimodal perception. *Advances in Neural Information Processing Systems*, 36:79142–79154.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. 2023. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 23555–23564.

Xiang Chen, Paul Pu Liang, Xuan Wang, Amir Zadeh, and Louis-Philippe Morency. 2023. Multiviz: Towards visualizing and understanding multimodal models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yong Dai, Duyu Tang, Liangxin Liu, Minghuan Tan, Cong Zhou, Jingquan Wang, Zhangyin Feng, Fan Zhang, Xueyu Hu, and Shuming Shi. 2022. One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. *arXiv preprint arXiv:2205.06126*.

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh,

[4] https://physionet.org/content/mimiciv/2.2/

and Guang Yang. 2024. Dynamic multimodal information bottleneck for multimodality classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7696–7706.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shi Feng, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2022. Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10135–10145.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, et al. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Wei Han, Hui Chen, and Soujanya Poria. 2021. 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*, pages 9180–9192.

Xing Han, Huy Nguyen, Carl Harris, Nhat Ho, and Suchi Saria. 2024. Fusemoe: Mixture-of-experts transformers for fleximodal fusion. In *Advances in Neural Information Processing Systems*, volume 37, pages 67850–67900. Curran Associates, Inc.

Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. 2024. Efficient modality selection in multimodal learning. *Journal of Machine Learning Research*, 25(47):1–39.

Jun Hou and Lucy Lu Wang. 2025. Explainable ai for clinical outcome prediction: a survey of clinician perceptions and preferences. *AMIA Summits on Translational Science Proceedings*, 2025:215.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Konstantinos Kontras, Thomas Strypsteen, Christos Chatzichristos, Paul Pu Liang, Matthew Blaschko, and Maarten De Vos. 2024. Multimodal fusion balancing through game-theoretic regularization. *arXiv preprint arXiv:2411.07335*.

Solomon Kullback. 1997. *Information theory and statistics*. Courier Corporation.

Yunxin Li, Shenyuan Jiang, Baotian Hu, Longyue Wang, Wanqi Zhong, Wenhan Luo, Lin Ma, and Min Zhang. 2025. Uni-moe: Scaling unified multimodal llms with mixture of experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, et al. 2024a. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947*.

Ronghao Lin and Haifeng Hu. 2023. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702.

Ronghao Lin and Haifeng Hu. 2024. Adapt and explore: Multimodal mixup for representation learning. *Information Fusion*, 105:102216.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. 2024b. Moma: Efficient early-fusion pre-training with mixture of modality-aware experts. *arXiv preprint arXiv:2407.21770*.

Yiwei Lyu, Paul Pu Liang, Zihao Deng, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Dime: Fine-grained interpretations of multimodal models via disentangled local explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 455–467.

Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. 2022. Multimodal contrastive learning with limoe: the language-image mixture of experts. *Advances in Neural Information Processing Systems*, 35:9564–9576.

Petra Poklukar, Miguel Vasco, Hang Yin, Francisco S Melo, Ana Paiva, and Danica Kragic. 2022. Geometric multimodal contrastive representation learning. In *International Conference on Machine Learning*, pages 17782–17800. PMLR.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. 2022. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149.

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

Betty van Aken, Jens-Michalis Papaioannou, Manuel Mayrdorfer, Klemens Budde, Felix A. Gers, and Alexander Löser. 2021. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 881–893. Association for Computational Linguistics.

Paul L Williams and Randall D Beer. 2010. Non-negative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Shaoxiang Wu, Damai Dai, Ziwei Qin, Tianyu Liu, Binghuai Lin, Yunbo Cao, and Zhifang Sui. 2023. Denoising bottleneck with mutual information maximization for video multimodal fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2231–2243.

Zihui Xue and Radu Marculescu. 2023. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2575–2584.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *Advances in Neural Information Processing Systems*, volume 37, pages 98782–98805. Curran Associates, Inc.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Yunhua Zhang, Hazel Doughty, and Cees Snoek. 2023. Learning unseen modality interaction. *Advances in Neural Information Processing Systems*, 36:54716–54726.

Xueliang Zhao, Mingyang Wang, Yingchun Tan, and Xianjie Wang. 2024. Tgmoe: A text guided mixture-of-experts model for multimodal sentiment analysis. *International Journal of Advanced Computer Science & Applications*, 15(8).

A MIMIC-IV with Missing Modalities

In real-world, the more modalities included will result in more likely and challenging missing modality problem. We conducted additional experiments using the MIMIC-IV dataset including stays with missing modalities as shown in Table 7. While the BTW-local (KL) variant achieves slightly better Macro-F1 (39.67%), we observe a performance drop after applying the full BTW weights, particularly when incorporating modality-level MI. These results indicate instance-level weights provide robust and stable performance even when modalities are absent. In contrast, the modality-level MI becomes unreliable, since the missing modality contributes no information, fooled the weights wrongly favors modalities that are consistently present regardless of their quality in informativeness. This bias degrades the variance stabilization ability for the bi-level weights, revealing the inherent sensitivity of MI-based alignment to incomplete real-world data.

Table 7: Performance on the MIMIC-IV dataset with randomly missing modalities. BTW-local (KL) and BTW represent variants of our bi-level weighting framework. All metrics are averaged over three random seeds. Standard deviations are shown in ±.

Method	Accuracy	Macro-F1	Weighted-F1
MulT	46.67±1.53	38.33±3.06	45.67±1.53
FuseMoE	44.00±1.00	38.67±1.53	44.00±2.00
BTW-local (KL)	45.67±0.58	39.67±1.15	45.00±0.00
BTW	44.00±1.41	35.50±0.71	42.50±0.71

B 3-Modalities Performance

This section presents additional results evaluating the robustness of our bi-level weighting framework against FuseMoE (Han et al., 2024) under various modality ablation scenarios on MIMIC-IV (Johnson et al., 2023). We compare model performance when each of the three auxiliary modalities (ECG, CXR, Text) is removed individually, under both no missing (Table 8 and full datasets Table 9).

Table 8: Performance comparison across weighting strategies on MIMIC without missing modalities

Method	Modality	Acc	Macro F1	Weighted F1
FuseMoE	w/o ECG	46	41	46
	w/o Text	43	42	41
	w/o CXR	48	37	47
BTW-local (KL)	w/o ECG	47	40	46
	w/o Text	42	33.5	40
	w/o CXR	44	38.5	44
BTW	w/o ECG	45	37	44
	w/o Text	45	17	29
	w/o CXR	46	36	43

Table 9: Performance comparison across weighting strategies on full MIMIC dataset

Method	Modality	Acc	Macro F1	Weighted F1
FuseMoE	w/o ECG	45	38	43
	w/o Text	43	35	40
	w/o CXR	47	38	46
BTW-local (KL)	w/o ECG	46	36	44
	w/o Text	44	30	40
	w/o CXR	44	36	42
BTW	w/o ECG	46	34	43
	w/o Text	44	25	36
	w/o CXR	46	37	45

C Proof of Estimation of Conditional Variance

This section formally derives the instance-level KL divergence used in our weighting framework for regression tasks. We provide a justification based on the Law of Total Variance, show the unbiasedness of residuals, and present the closed-form expression for Gaussian KL divergence between unimodal and multimodal predictions.

C.1 Law of Total Variance

$$\mathrm{Var}(Y) = \mathrm{Var}\big(\mathrm{E}[Y \mid X]\big) + \mathrm{E}\big[\mathrm{Var}(Y \mid X)\big] \tag{11}$$

$$\mathrm{Var}\big(\mathrm{E}[Y \mid X_i]\big) = \mathrm{Var}(\mu_i) = 0 \tag{12}$$

$$\mathrm{Var}(Y \mid X_i) = \mathrm{Var}(Y) = \mathrm{E}\big[(Y - \mu_i)^2\big] \tag{13}$$

C.2 Unbiasedness of the Squared Residual

$$\mathrm{E}\big[(Y - \mu_i)^2 \mid X_i\big] = \mathrm{Var}(Y \mid X_i), \tag{14}$$

so each empirical squared residual $(Y_i - \mu_i)^2$ is an unbiased estimate of the conditional variance.

C.3 Gaussian KL–Divergence

For two Gaussians $p = \mathcal{N}(\mu_p, \sigma_p)$ and $q = \mathcal{N}(\mu_q, \sigma_q)$,

$$D_{\mathrm{KL}}(p \parallel q) = \log\frac{\sigma_q}{\sigma_p} + \frac{\sigma_p^2 + (\mu_p - \mu_q)^2}{2\,\sigma_q^2} - \tfrac{1}{2}. \tag{15}$$

By setting

$$\mu_p = \mu_i, \quad \sigma_p^2 = (Y_i - \mu_i)^2,$$

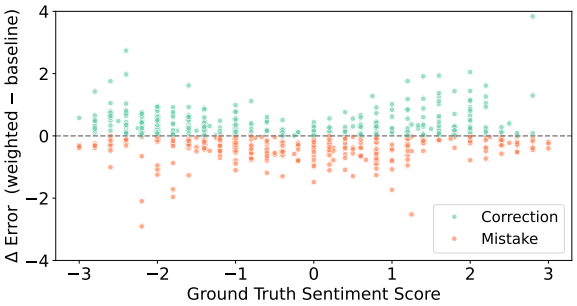
$$\mu_q = \mu_i^{\text{multi}}, \quad \sigma_q = (Y_i - \mu_i^{\text{multi}})^2,$$

we obtain the instance–level weight

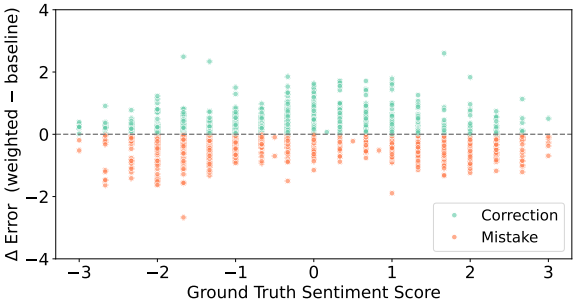
$$w_i = D_{\text{KL}}\big(\mathcal{N}(\mu_i, (Y_i - \mu_i)^2) \,\|\, \mathcal{N}(\mu_i^{\text{multi}}, (Y_i - \mu_i^{\text{multi}})^2\big). \quad (16)$$

D Scatter plots for error analysis

We visualize instance-level corrections and mistakes for sentiment regression as shown in Figure 6 and Figure 7 for BTW-local (KL) and BTW scenario, respectively. Each point represents the change in prediction (corrected or error) relative to the true sentiment score, with the vertical axis indicating the magnitude and sign of these modifications. For MOSI, corrections using BTW-local (KL) tend to show larger error reductions in moderately positive and negative regions, whereas the BTW approach produces a denser band of modest corrections across the full range. In MOSEI, both schemes demonstrate more uniform correction magnitudes across sentiment values. In all cases, the majority of mistakes (orange) are smaller in magnitude and more evenly distributed, confirming the effectiveness of bi-level weights.

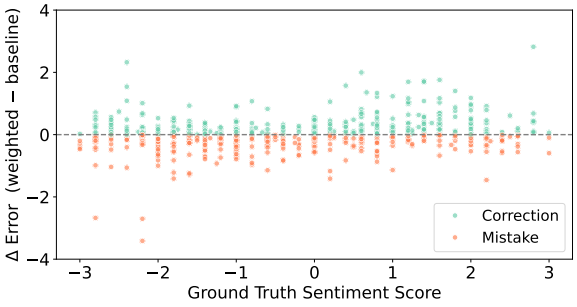


(a) MOSI BTW-local (KL) Corrections and Mistakes Distribution

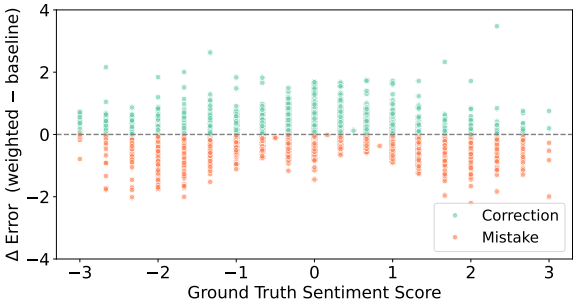


(b) MOSEI BTW-local (KL) Corrections and Mistakes Distribution

Figure 6: Visualization of test split scatter plots, corrections density and mistakes density over the ground truth sentiment score for BTW-local (KL) experiments.



(a) MOSI BTW Corrections and Mistakes Distribution



(b) MOSEI BTW Corrections and Mistakes Distribution

Figure 7: Visualization of test split scatter plots corrections density and mistakes density over the ground truth sentiment score for BTW experiments.

E Proof of Mutual Information

Let $\mathcal{X}$ be the whole dataset with N data points. Naturally $P(X_i) = 1/N$ for all $X_i \in \mathcal{X}$. Then we have:

$$P(\hat{y}^{(m)}) = \sum_{X_i^{(m)} \in \mathcal{X}} P(\hat{y}_i^{(m)} | X_i^{(m)})\, P(X_i^{(m)}) \quad (17)$$

$$P(\hat{y}^{\text{multi}}) = \sum_{X_i^{\text{multi}} \in \mathcal{X}} P(\hat{y}_i^{\text{multi}} | X_i^{\text{multi}})\, P(X_i^{\text{multi}}) \quad (18)$$

And the joint probability is defined as:

$$P(\hat{y}^{(m)}, \hat{y}^{\text{multi}}) = \sum_{X_i^{(m)} \in \mathcal{X}} P(\hat{y}_i^{(m)} | X_i^{(m)}) P(X_i^{(m)} | X_i^{\text{multi}}) P(\hat{y}_i^{\text{multi}} | X_i^{\text{multi}}) P(X_i^{\text{multi}}) \quad (19)$$

F Computation Resources, Time Cost, and Hyper-Parameters

We summarize the computation resources and MoE-specific hyperparameters and packages used in our experiments. Hyper-parameter settings are

Table 10: Computational cost analysis. All times are averaged over three runs.

	Method	Text Overhead	Audio Overhead	Video Overhead	Main Training Time	Total Est. Time
MOSI	Baseline	N/A	N/A	N/A	6m 45s	6m 45s
	BTW-local (KL)	6m 4s	2m 39s	1m 7s	7m 35s	17m 25s
	BTW (KL+MI)	6m 4s	2m 39s	1m 7s	7m 55s	17m 46s
MOSEI	Baseline	N/A	N/A	N/A	37m 36s	37m 36s
	BTW-local (KL)	30m 4s	22m 24s	22m 20s	43m 59s	118m 47s
	BTW (KL+MI)	30m 4s	22m 24s	22m 20s	45m 36s	120m 24s

Table 11: Hyperparameters used for MoE module

Parameter Name	Value
Number of Experts	16
FFN hidden size	512
Top k	2
Router_type	permod
Hidden activation function	GeLU
Number of MoE layers	3

aligned with those in prior work on LOS classification and sentiment analysis to ensure comparability and reproducibility.

Computation Resources All experiments were conducted on a single NVIDIA A40 GPU with 48GB memory (CUDA 12.4, driver version 550.67). For the CMU datasets, we follow the preprocessing steps from MMIM (Han et al., 2021) and incorporate the MoE layer from FuseMoE (Han et al., 2024).

Time Cost Our framework introduces an initial, one-time overhead for pre-training the unimodal models, which is required to initialize the weighting process. The duration of this step is highly dependent on the size of the dataset. As shown in Table 10, the per-epoch training time of our methods remains comparable to the baseline.

Hyper-Parameter Settings The hyper-parameters for MoE module are listed in Table 11. Other hyper-parameters used for LOS classification are same as in (Han et al., 2024), and for sentiment analysis are same as in (Han et al., 2021).

Implementation Our implementation used Python 3.8 with the following key libraries: PyTorch, NumPy, pandas, scikit-learn, and HuggingFace Transformers.