Online Supplement

When Multimodal Interactions Impair Prediction: A Novel Regularized Deep Learning Strategy

Gang Chen, Shuaiyong Xiao, Chenghong Zhang, Huimin Zhao

Appendix A. Exploratory Analysis Results

A.1 Ablation Analysis

The ablation of NIRMD yielded five ablated variants: NIRMD-FS (i.e., $\lambda_1 = 0$, without feature sparsity), NIRMD-NEMII (i.e., $\lambda_2 = 0$, without NEMII regularization), NIRMD-NEFI (i.e., $\alpha = 0$, without NEFI regularization), NIRMD-NEMI (i.e., $\alpha = 1$, without NEMI regularization), and NIRMD-NEFI-NEMI (i.e., without both NEFI and NEMI regularizers). Table A1 summarizes the ablation experiment results. NIRMD stably outperformed all ablated variants, demonstrating the indispensable roles of all components (Table 3) in better leveraging multimodal data for prediction. All ablated variants of NIRMD stably outperformed their corresponding benchmarks, corroborating the value and rationality of our design rationales (Table 3) for tackling multimodal interactions. Especially, ablating the L_1-norm-based regularizer (i.e., NIRMD-FS) decreased the benchmarks' prediction performance, indicating that issues arising from the curse of dimensionality (e.g., noise and irrelevant features) are also present in the multimodal deep learning context. Moreover, ablating the regularizers we designed specifically to mitigate negative multimodal interactions resulted in further decline in prediction performance, suggesting that, in addition to the curse of dimensionality, multimodal deep learning faces challenges from negative multimodal interactions.

Table A1. Mean (Standard Deviation) of Prediction Performance of NIRMD vs Ablated Variants.

Case	Method	Without NIRMD or Variant	With Ablated Variant					With NIRMD
			NIRMD-FS	NIRMD-NEMII	NIRMD-NEFI	NIRMD-NEMI	NIRMD-NEFI-NEMI	
SVAP (RMSE%)	MultiEMO	33.16 (4.27)	32.40 (1.51)	29.52 (1.09)	31.55 (1.16)	25.43 (0.56)	32.23 (1.49)	**23.60 (1.37)**
	DMVAE	33.33 (0.83)	32.98 (1.16)	31.71 (0.91)	32.32 (1.23)	31.08 (0.93)	32.52 (1.47)	**28.34 (0.80)**
	MFIR	29.52 (2.15)	28.78 (1.06)	25.56 (0.97)	26.19 (1.03)	25.64 (0.90)	29.40 (1.06)	**22.51 (0.63)**
	MMoE	28.56 (3.56)	27.32 (4.29)	27.14 (4.63)	26.76 (3.85)	26.36 (1.80)	28.34 (4.81)	**22.99 (0.84)**
MRSC (AUC%)	MultiEMO	85.15 (2.68)	88.29 (3.47)	89.14 (3.49)	86.35 (5.06)	88.28 (3.52)	85.98 (3.83)	**92.41 (1.97)**
	TCGAN	87.83 (1.16)	88.73 (0.83)	88.79 (0.89)	88.44 (1.02)	88.53 (0.91)	87.95 (1.11)	**89.96 (1.43)**
	MVCN	87.39 (2.02)	88.84 (2.72)	89.01 (3.01)	88.79 (2.61)	88.76 (2.71)	87.42 (2.71)	**92.98 (1.97)**
	MMoE	83.19 (3.76)	85.11 (4.88)	85.19 (6.41)	85.83 (5.46)	84.32 (5.49)	84.26 (5.05)	**91.52 (2.30)**
DRP (AUC%)	HGMF	83.12 (1.75)	83.41 (2.39)	83.60 (1.99)	83.39 (2.06)	83.89 (4.62)	83.29 (0.05)	**85.75 (2.07)**
	DMVAE	82.49 (1.84)	83.89 (1.63)	84.20 (2.09)	84.35 (1.72)	83.99 (1.52)	82.55 (3.41)	**85.68 (2.64)**
	MFIR	72.94 (2.55)	83.23 (2.95)	83.39 (3.15)	83.21 (2.62)	83.56 (2.79)	81.80 (3.20)	**86.98 (1.98)**
	MMoE	74.16 (2.93)	78.90 (3.21)	78.93 (4.29)	78.94 (3.66)	79.50 (3.40)	77.78 (3.96)	**83.78 (3.33)**

A.2 Contribution of NIRMD-Regularized Multimodal Representations

To verify the regularization effectiveness of NIRMD for multimodal interactions, we analyzed the marginal predictive contributions of multimodal representation features learned by the benchmarks with and without NIRMD, respectively, using the Shapley value (Aas et al. 2021). We selected the best performing methods, in terms of the multimodal interaction regularization effect, from each of the four classes of benchmarks for each of the three cases (Table 4). Specifically, in each case study, we obtained the multimodal representation features through the trained benchmark models with and without NIRMD, respectively. Then, we calculated the Shapley values of the representation features in each modality and summed them to determine the total predictive contribution of all multimodal representations. As shown in Tables A2 to A4, the predictive contributions of some modalities were negative, implying the presence of negative multimodal interactions that impair prediction. By virtue of NIRMD, the predictive contributions of multimodal representation features learned by the benchmark methods were consistently enhanced, demonstrating the effectiveness of NIRMD in encouraging positive multimodal interactions while mitigating negative ones.

Table A2. Predictive Contributions of NIRMD-Regularized Multimodal Representations in the SVAP Case.

Method	Q	T	G	A	V	Total
MultiEMO	-0.1517	0.2021	0.1800	0.4612	0.1286	0.8203
MultiEMO+NIRMD	0.1689	0.1620	0.1694	0.3474	0.1483	**0.9960**
DMVAE	-0.1257	0.1703	0.1754	0.1594	0.1415	0.5209
DMVAE+NIRMD	0.0810	0.2213	0.1374	0.2991	0.2437	**0.9825**
MFIR	-0.2224	0.1716	0.1666	0.1539	0.1438	0.4135
MFIR+NIRMD	0.1374	0.1718	0.1739	0.1591	0.1320	**0.7743**
MMoE	-0.1374	0.1612	0.1773	0.1598	0.1421	0.5029
MMoE+NIRMD	0.1263	0.1465	0.1938	0.2979	0.1868	**0.9513**

Table A3. Predictive Contributions of NIRMD-Regularized Multimodal Representations in the MRSC Case.

Method	Q	T	G	V	Total
MultiEMO	0.0272	0.2719	-0.0082	-0.0015	0.2894
MultiEMO+NIRMD	0.0276	0.4777	0.0001	0.0024	**0.5078**
TCGAN	0.0293	0.2807	-0.0146	-0.0025	0.2929
TCGAN+NIRMD	0.0467	0.4598	0.0007	0.0036	**0.5109**
MVCN	0.0158	0.2591	-0.0177	0.0024	0.2596
MVCN+NIRMD	0.0241	0.4680	0.0135	0.0044	**0.5101**
MMoE	0.0268	0.2412	-0.0127	-0.0017	0.2535
MMoE+NIRMD	0.0329	0.4468	0.0136	0.0034	**0.4967**

Table A4. Predictive Contributions of NIRMD-Regularized Multimodal Representations in the DRP Case.

Method	H	S	T	G	Total
HGMF	0.0834	0.0181	-0.0194	0.0200	0.1022
HGMF+NIRMD	0.1107	0.0435	0.0151	0.0619	**0.2312**

DMVAE	0.0155	0.0105	-0.0254	0.0140	0.0146
DMVAE+NIRMD	0.0358	0.0079	0.0001	0.0231	**0.0668**
MFIR	0.1080	0.0082	0.0743	-0.0098	0.1807
MFIR+NIRMD	0.0640	0.4230	0.0084	0.0005	**0.4959**
MMoE	0.1069	0.0100	0.0600	-0.0154	0.1615
MMoE+NIRMD	0.0714	0.3036	0.0048	0.0118	**0.3916**

A.3 Case Analysis of NIRMD's Regularization Effect on Negative Multimodal Interactions

Table A5 shows the multimodal prediction performance of the benchmark method MVCN with and without NIRMD, respectively, in the SVAP case. Figure A1 provides a visual explanation of the contributions of multimodal features leveraged by MVCN, with and without NIRMD, respectively, on a real instance from the SVAP case for prediction, using Grad-CAM (Selvaraju et al. 2017). As shown in Table A5, the interaction among image features (i.e., feature-level interaction) and the interaction between the textual and visual modalities (i.e., modality-level interaction) impaired prediction, confirming the presence of negative multimodal interactions. Besides, as shown in Figure A1, NIRMD enabled MVCN to focus more effectively on positively interacting features that are relevant and mutually complementary in revealing the hotel's attractiveness, demonstrating the effectiveness of NIRMD in learning positively interacting multimodal representations to enhance prediction.

Table A5. Multimodal Prediction Performance of MVCN with vs without NIRMD Based on the Example Instance in the SVAP Case.

Method	Feature								Fusion Best
	Frames 4-5		Frames 4-5+Text		Frames 1-5		All Frames+Text		
	Prediction Score	Prediction Error	Prediction Score	Prediction Error	Prediction Score	Prediction Error	Prediction Score	Prediction Error	
MVCN	12.32	0.58	12.43	0.47	14.36	1.46	16.94	4.04	×
MVCN+ NIRMD	12.60	0.30	13.25	0.35	13.33	0.43	13.19	0.29	√

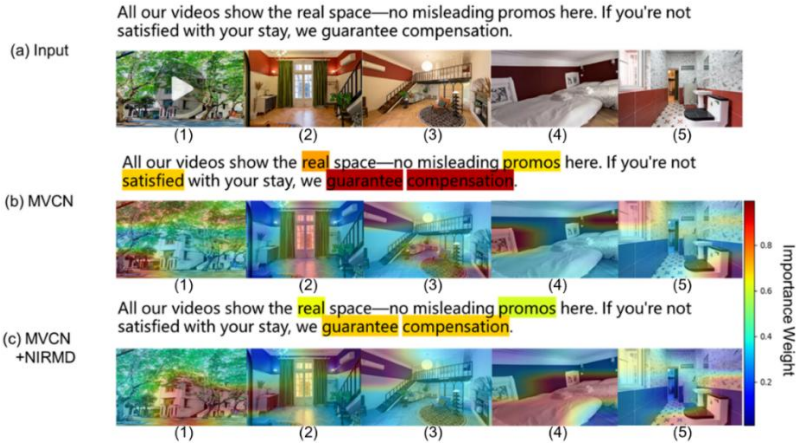

Figure A1. Visual Explanation of Multimodal Features' Predictive Contributions Learned by MVCN with vs without NIRMD.

Appendix B. Theorem Proofs

B.1 Proof of Theorem 1

Suppose η_j is a positive constant that satisfies the condition $\left|\left(1-\frac{\left(x_i^R-x_j^I\right)^2}{2n}\right)\text{sign}(w_i)\right| \le 1-\eta_j$, where $\boldsymbol{x}_i^R \in \boldsymbol{X}^R$ (i.e., relevant features) and $\boldsymbol{x}_j^I \in \boldsymbol{X}^I$ (i.e., irrelevant features). Then, we have

$$p\big(\text{sign}(\widehat{\boldsymbol{w}}) = \text{sign}(\boldsymbol{w})\big) \ge p(A \cap B), \tag{B.1}$$

according to Zhao and Yu (2006), where $\text{sign}(\boldsymbol{w})$ signifies the consistent selection of the true relevant features for prediction, and $\text{sign}(\widehat{\boldsymbol{w}})$ denotes the estimated model's selection of relevant features, which can be achieved by the regularizer $\|\boldsymbol{w}\|_1$. $A = \left\{|\chi_i| \le \sqrt{n}\left(|w_i| - \frac{\lambda}{2n}\varpi_i\right), \forall \boldsymbol{x}_i^R \in \boldsymbol{X}^R\right\}$ and $B = \left\{\left|\gamma_j\right| \le \frac{\lambda}{2\sqrt{n}}\eta_j, \forall \boldsymbol{x}_i^R \in \boldsymbol{X}^R, \boldsymbol{x}_j^I \in \boldsymbol{X}^I\right\}$, where $\chi_i = \frac{\left(\boldsymbol{x}_i^R\right)^\top \boldsymbol{\varepsilon}}{\sqrt{n}}$, $\gamma_j = \left(1 - \frac{\left(x_i^R-x_j^I\right)^2}{2n}\right)\frac{\left(\boldsymbol{x}_i^R\right)^\top \boldsymbol{\varepsilon}}{\sqrt{n}} - \frac{\left(\boldsymbol{x}_j^I\right)^\top \boldsymbol{\varepsilon}}{\sqrt{n}}$, and $\varpi_i = \text{sign}(w_i)$. ε denotes the error term.

Given that $p(A \cup B) \le 1$, we can write $p(A \cup B) = p(A) + p(B) - p(A \cap B) \le 1$. Let

$$p\big(\hat{A}\big) = \sum_{i=1}^R p\left(|\chi_i| \ge \sqrt{n}\left(|w_i| - \frac{\lambda}{2n}\varpi_i\right)\right), \tag{B.2}$$

$$p\big(\hat{B}\big) = \sum_{j=1}^I p\left(\left|\gamma_j\right| \ge \frac{\lambda}{2\sqrt{n}}\eta_j\right), \tag{B.3}$$

we have $p(A) = 1 - p\big(\hat{A}\big)$ and $p(B) = 1 - p\big(\hat{B}\big)$, and further, $p(A \cap B) \ge 1 - \left(p\big(\hat{A}\big) + p\big(\hat{B}\big)\right)$. Hence, we have

$$p\big(\text{sign}(\widehat{\boldsymbol{w}}) = \text{sign}(\boldsymbol{w})\big) \ge p(A \cap B) \ge 1 - \left(p\big(\hat{A}\big) + p\big(\hat{B}\big)\right). \tag{A.4}$$

To enhance the probability $p\big(\text{sign}(\widehat{\boldsymbol{w}}) = \text{sign}(\boldsymbol{w})\big)$, it suffices to minimize $p\big(\hat{A}\big) + p\big(\hat{B}\big)$. According to Knight and Fu (2000), as the instance size n increases,

$$\chi_i \xrightarrow{d} \mathcal{N}(0,\ 1), \tag{B.5}$$

$$\gamma_j \xrightarrow{d} \mathcal{N}\big(0, 1 - \rho_{ij}^2\big), \tag{B.6}$$

where ρ_{ij} represents the correlation between $\boldsymbol{x}_i^R$ and $\boldsymbol{x}_j^I$.

Applying Mill's inequality, we get the following upper bounds for $p(\hat{A})$ and $p(\hat{B})$:

$$p\big(\hat{A}\big) \le \sum_{i=1}^R p\left(1 - \Phi\left(\sqrt{n}\left(|w_i| - \frac{\lambda}{2n}\varpi_i\right)\right)\right) = o\big(e^{-n^c}\big), \tag{B.7}$$

$$p\big(\hat{B}\big) \le \sum_{j=1}^I p\left(1 - \Phi\left(\frac{\lambda}{2\sqrt{n\big(1-\rho_{ij}^2\big)}}\eta_j\right)\right) = o\big(e^{-n^c}\big), \tag{B.8}$$

where $0 < c < 1$ is a constant.

Combining these upper bounds for $p(\hat{A})$ and $p(\hat{B})$, we obtain

$$p(\hat{A}) + p(\hat{B}) \leq o(e^{-n^c}) + o(e^{-n^c}) = o(e^{-n^c}), \tag{B.9}$$

and consequently,

$$p(A \cap B) \geq 1 - o(e^{-n^c}). \tag{B.10}$$

Thus,

$$p(\text{sign}(\widehat{\boldsymbol{w}}) = \text{sign}(\mathbf{w})) \geq 1 - o(e^{-n^c}), \tag{B.11}$$

indicating that enlarging the regularizer $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ decreases $\left(1 - \frac{(x_i^R - x_j^I)^2}{2n}\right)$, which in turn increases $p(\text{sign}(\widehat{\boldsymbol{w}}) = \text{sign}(\mathbf{w}))$. That is, more relevant features will be identified and learned.

In summary, enlarging the regularizer $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$ not only decorrelates $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, i.e., $I(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{y})$ decreases, but also strengthens the learning (selection) of relevant features, i.e., $I(\boldsymbol{x}_j, \boldsymbol{y}|\boldsymbol{x}_i)$ and $I(\boldsymbol{x}_i, \boldsymbol{y}|\boldsymbol{x}_j)$ increase. Based on the definition of NEFI in Definition 3.1, we deduce that the regularizers $\{\|\boldsymbol{w}\|_1 - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\}$ can eliminate the predictive influence of negatively interacting features with a certain probability, i.e., $p(sign[\widehat{\boldsymbol{w}}^P, \widehat{\boldsymbol{w}}^N] = sign[\boldsymbol{w}^P, \mathbf{0}])$ $\xrightarrow{Regularizer\{\|\boldsymbol{w}\|_1 - \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2\}}$ $1 - o(e^{-n^c})$, as the size of instances enlarges. Consequently, NEFI can be effectively mitigated by the inclusion of these two regularizers.

B.2 Proof of Theorem 2

Consider weighting modality-wise instances with learnable parameters $\boldsymbol{U}$, where $0 \leq u_i^{(m)} \in \boldsymbol{U} \leq 1$, through the NEMII regularizer. The estimator $\widehat{\boldsymbol{w}}_{NEMII}$ can be expressed as

$$\widehat{\boldsymbol{w}}_{NEMII} = argmin_{\boldsymbol{w}} \sum_{i=1}^{b} u_i (y_i - x_i w_i)^2. \tag{B.12}$$

The closed-form solution for the $\widehat{\boldsymbol{w}}_{NEMII}$ estimator is

$$\widehat{\boldsymbol{w}}_{NEMII} = (\boldsymbol{X}^\top \boldsymbol{U} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{U} \boldsymbol{y}, \tag{B.13}$$

whereas the solution for $\widehat{\boldsymbol{w}}$ (i.e., the feature weight estimator without the NEMII regularizer) is

$$\widehat{\boldsymbol{w}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{B.14}$$

The variance of the $\widehat{\boldsymbol{w}}_{NEMII}$ estimator is

$$\begin{aligned} Var(\widehat{\boldsymbol{w}}_{NEMII}) &= (\boldsymbol{X}^\top \boldsymbol{U} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{U} Var(\boldsymbol{\varepsilon}) \boldsymbol{U}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{U} \boldsymbol{X})^{-1} \\ &= \|\boldsymbol{\varepsilon}\|_2^2 (\boldsymbol{X}^\top \boldsymbol{U} \boldsymbol{X})^{-1}, \end{aligned} \tag{B.15}$$

where $Var(\boldsymbol{\varepsilon})$ represents the variances of the error terms.

In comparison, the variance of the $\widehat{\boldsymbol{w}}$ estimator is

$$Var(\widehat{\boldsymbol{w}}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top Var(\boldsymbol{\varepsilon}) \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}$$

$$= \|\boldsymbol{\varepsilon}\|_2^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \quad \text{(B.16)}$$

Since $0 \leq u \leq 1$, we have

$$Var(\widehat{\boldsymbol{w}}_{NEMII}) \leq Var(\widehat{\boldsymbol{w}}). \quad \text{(B.17)}$$

Consequently, by introducing the NEMII regularizer, the weights of multimodal representations will be learned more accurately, e.g., eliminating adverse modality-wise instances (i.e., $u_i^{(m)} = 0$) and downplaying the role of less important modality-wise instances during the training phase (i.e., $u_i^{(m)} < 1$), resulting in enhanced multimodal prediction performance.

Appendix C. The Training Algorithm of NIRMD

Parameters: $\lambda_1, \lambda_2, \lambda_3, \alpha, b$, E (number of epochs).
Input: [**D**, **y**]

Construct M feature modalities suitable for the M data modalities of the dataset [**D**, **y**].
// Fix **u** with instances in **D**

for epoch←1 to E
  initialize **u**.
  for *batch*←1 to b
    for m←1 to M
      Run a feedforward pass through the m-th pre-trained neural network to get its output representation feature matrix $\boldsymbol{X}^{(m)}$.
      Conduct batch normalization for $\boldsymbol{X}^{(m)}$.
    Concatenate $[\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(m)}, \cdots, \boldsymbol{X}^{(M)}]$
    Compute the total loss with Equation (7).
    //$\lambda_1 \|\boldsymbol{w}\|_1$ for facilitating feature relevance.
    // $\lambda_2 \sum_{m=1}^{M} \left\|\boldsymbol{U}^{(m)}\right\|_{2,1}$ for dealing with modality-wise instance-level interaction.
    //$\lambda_3 \alpha (\sum_{i<j} \left\|\boldsymbol{x}_i^{(m)} - \bar{\boldsymbol{x}}\right\|_2^2)$ for dealing with feature-level interaction.
    //$\lambda_3 (1-\alpha) (\sum_{k<q}^{M} \xi_{kq} \left\|\boldsymbol{X}^{(k)} - \boldsymbol{X}^{(q)}\right\|_F^2)$ for dealing with modality-level interaction.
    //α and $1-\alpha$ for reconciling multimodal interactions.
    // Loss backward propagation as follows.
    if concatenation layer
      Update **w** with mini-batch proximal gradient descent (MPGD). // Realizing trainable feature sparsity.
      Update **u** with MPGD. // Realizing trainable modality-wise instance sparsity.
    else
      Update **w** with mini-batch gradient descent.

Figure C1. The Training Algorithm of NIRMD.

Appendix D. Feature Modality Construction

Based on the collected multimodal data, we extracted multimodal features for each case using suitable neural networks. Table D1 presents the quantitative features we derived for constructing the quantitative feature modality.

Table D1. Quantitative Features Extracted.

Case	Quantitative Features
(1) SVAP	Text length (number of words), video duration, number of product types attached to the short video, number of fans (by the short video's release date), and number of followees.

(2) MRSC	Review length (number of words), number of accompanying images, number of replies to the review, and number of votes received by the review.
(3) DRP (hard features)	**Demographics**: gender, age, education level, marital status (married or not), number of children, income level, and housing condition (with a house or not); **Loan features**: loan amount, loan term, interest rate, prior loan history, mortgage, and loan purpose (for emergency or not).

In the SVAP case, we constructed five feature modalities for each sample (short video): quantitative feature modality (Q) based on video-related indicators, acoustic feature modality (A) based on audio data derived from the video, textual feature modality (T) based on text from transcribing the audio, semantic graph feature modality (G) constructed based on the text data, and visual temporal feature modality (V) based on image stream sampled from the video. In the MRSC case, we constructed four feature modalities for each sample (multimodal review): quantitative feature modality (Q) based on multimodal review-related indicators, textual feature modality (T) based on review text, semantic graph feature modality (G) constructed based on the review text, and visual feature modality (V) based on review images. In the DRP case, we constructed four feature modalities for each sample (borrower): hard feature modality (H) based on quantitative hard features, social interaction feature modality (S) based on quantitative social interaction indicators, textual feature modality (T) based on social media texts, and semantic graph feature modality (G) constructed based on the merged social media text document. Apparently, the textual feature modality and semantic graph feature modality may overlap to some extent as they are constructed based on the same raw data, presenting an opportunity for NIRMD to demonstrate its learning capability in challenging situations. We transcribed audio to text by calling a widely used commercial application program interface available at https://ai.unisound.com/?channel=unisound, which has been shown to be competent for various Chinese voice-to-text tasks. We preprocessed each video into an image stream using the pre-trained 3D-ResNet model (Hara et al. 2018).

D.1 Quantitative Feature Modality

To effectively leverage quantitative features for the given multimodal prediction task and better coordinate them with other feature modalities, we constructed the quantitative feature modality via an MLP. During the subsequent multimodal deep learning, where multiple feature modalities were synergistically trained while supervised by the target variable, the MLP was optimized. In so doing, the MLP can output deep representation features for its input quantitative features.

D.2 Textual Feature Modality

For extracting textual features, we used a pre-trained model BERT (Devlin et al. 2018) to embed language structures, semantic patterns, and linguistic styles of original texts into text embeddings.

Next, taking text embeddings as inputs, we adopted LSTM to construct a textual feature modality for multimodal prediction. Further, we used attention mechanism (Vaswani et al. 2017) based on LSTM to endow the textual feature modality with the ability to attend more to those salient representation features learned by LSTM. During the subsequent multimodal deep learning, the LSTM-attention-based textual feature modality was unified with other feature modalities toward the given multimodal prediction task.

D.3 Semantic Graph Feature Modality

Unlike LSTM, which encodes texts by taking them as word sequences, GCN (Liu et al. 2020) enables the semantic relationships among the keywords to be encoded into a network spatially (Yao et al. 2019). Through the graph convolution process based on the keywords-based network, the global semantic structures and the sequential-free narrative patterns in text data can be effectively exploited. To construct a semantic graph feature modality, we first used Term Frequency-Inverse Document Frequency (TF-IDF)-weighted n-gram (unigram) model for keyword extraction based on all texts of the samples. Then, we selected a number of top keywords from the candidate words according to their average TF-IDF values over all texts (samples). Next, for each sample, we constructed a semantic graph (represented as an adjacency matrix) using the selected keywords and their TF-IDF values. In the adjacency matrix, each row/column corresponds to one of the selected keywords, and each diagonal (non-diagonal) element is the keyword's TF-IDF value over the sample (difference between the two keywords' TF-IDF values). In this way, the adjacency matrix embeds the complex and global semantic structures of the text into a keyword space. We then fed the adjacency matrix coupled with the TF-IDF-based keyword features into a GCN to get a semantic graph feature modality, which, during the subsequent multimodal deep learning, was unified with other feature modalities toward multimodal prediction.

D.4 Acoustic Feature Modality

Audio information is valuable for video-based multimodal prediction tasks. For example, in the SVAP case, audio information constitutes the vocal attractiveness directly (Petty et al. 2020). To extract audio features, we adopted the mel-frequency cepstral coefficients (MFCC), a common technique for deriving a parametric representation of acoustic signals based on the human auditory system (i.e., with the frequency varying within the human ear's critical bandwidth) (Zheng et al. 2001). After that, we input the MFCC matrix into an LSTM-attention model to get the acoustic feature modality, which was unified with other feature modalities during the subsequent

multimodal deep learning.

D.5 Visual Feature Modality and Visual-Temporal Feature Modality

Two visual-related data modalities, i.e., images and image streams, commonly appear in multimodal prediction tasks. In the MRSC case, each multimodal review is associated with one or more images. The visual cues provided by review images have been shown to be valuable for assessing the sentiment implied in the review (Xu et al. 2019). To construct the visual feature modality, we put the image into a pre-trained VGG-16 neural network (Simonyan and Zisserman 2014). During the subsequent multimodal deep learning process, we added several new dense layers on the VGG-16 model to adapt it to multimodal prediction. Images sampled from a video essentially constitute an image stream. We adopted 3D-ResNet, a popular network for encoding video data into a sequential embedding with spatio-temporal (3D) kernels (Hara et al. 2018), to obtain sequential embeddings of the image streams. Based on such embeddings, we used an LSTM-attention model to construct the visual-temporal feature modality, which was subsequently coordinated with other feature modalities during the subsequent multimodal deep learning.

Appendix E. Parameter Configuration

In the experiments, we performed parameter tuning for the benchmarks and NIRMD based on each experiment setting to make them better fit the multimodal prediction tasks (i.e., SVAP, MRSC, and DRP). Table E1 summarizes the experiment settings of some key parameters.

Table E1. Parameter Setting in the Experiment.

Model	Parameter	Setting	Description
NIRMD	Epochs	5-10	Number of epochs to train the model
	Batch size	32-64	Number of samples per gradient update
	λ_1	Lambda values	Threshold factor of feature weights' L1-norm
	λ_2	Lambda values	Threshold factor of modality-wise instance weights' L21-norm
	λ_3	Lambda values	Threshold factor of NEFI and NEMI
	α	0.1 to 0.9 with an interval of 0.1	The trade-off factor of regularizations on NEFI and NEMI

Lambda values: 0.00001, 0.00005, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1.

Appendix F. Statistical Significance Test Results

Table F1. Tukey-Kramer Test Result (t and p) of NIRMD vs Benchmarks.

Method	SVAP (RMSE%)	MRSC (AUC%)	DRP (AUC%)
HGMF	54.84***	11.96***	9.61***
STAN	31.89***	15.90***	9.29***
MultiEMO	16.44***	21.84***	25.91***
RJCMA	11.38***	15.73***	19.61***
DMVAE	42.85***	17.11***	9.81***
TCGAN	97.69***	11.45***	6.90***
DCCA	33.68***	18.52***	7.20***
MVCN	22.93***	19.80***	65.00***
Dropout	4.67**	6.33***	12.83***

AWD	8.63***	5.42***	17.97***
HLGM	4.88***	5.86***	26.51***
SLMV	97.26***	19.21***	46.91***
MFIR	24.37***	18.67***	43.46***
MoEF	5.70***	20.54***	19.24***
MMoE	10.14***	18.91***	21.64***

*** Significant at the 0.001 level; ** Significant at the 0.01 level; * Significant at the 0.05 level.

Table F2. Tukey-Kramer Test Result (t and p) of NIRMD vs Ablated Variants.

Case	Method	Ablated Variant				
		NIRMD-FS	NIRMD-NEMII	NIRMD-NEFI	NIRMD-NEMI	NIRMD-NEFI-NEMI
SVAP (RMSE%)	MultiEMO+NIRMD	37.97***	21.81***	33.22***	1.09	29.57***
	DMVAE+NIRMD	31.88***	23.15***	27.34***	18.82***	24.98***
	MFIR+NIRMD	35.82***	11.30***	16.12***	11.93***	43.54***
	MMoE+NIRMD	5.26**	4.87**	4.11*	3.29	7.23***
MRSC (AUC%)	MultiEMO+NIRMD	7.99***	6.34***	11.81***	8.01***	14.94***
	TCGAN+NIRMD	8.37***	7.96***	10.37***	9.72***	11.09***
	MVCN+NIRMD	11.12***	10.68***	11.27***	11.35***	16.58***
	MMoE+NIRMD	8.89***	8.78***	7.89***	9.99***	13.07***
DRP (AUC%)	HGMF+NIRMD	5.87***	5.41**	5.93***	4.67**	11.87***
	DMVAE+NIRMD	9.71***	9.28***	9.76***	8.86***	7.25***
	MFIR+NIRMD	6.38***	5.28**	4.74**	6.03***	13.76***
	MMoE+NIRMD	9.58***	9.51***	9.49***	8.39***	11.58***

*** Significant at the 0.001 level;** Significant at the 0.01 level;* Significant at the 0.05 level.

Appendix G. Computational Complexity Analysis

Figure G1 shows the time complexity of the benchmarks with vs without NIRMD in the three cases, where the horizontal axis represents the benchmark category and the vertical axis shows the number of instances processed per second during the training phase. Table G1 presents the space complexity (i.e., the number of parameters) of the benchmarks with vs without NIRMD in the three cases. As shown in Figure G1 and Table G1, incorporating NIRMD did not lead to a significant increase in computational complexity, demonstrating its potential for widespread deployment and practical application.

Table G1. Space Complexity of the Benchmarks with vs without NIRMD.

Method	SVAP		MRSC		DRP	
	Benchmark	Benchmark+NIRMD	Benchmark	Benchmark+NIRMD	Benchmark	Benchmark+NIRMD
HGMF	522,977	572,987 (9.56%)	379,145	391,151 (3.17%)	240,561	269,187 (11.90%)
STAN	552,643	602,653 (9.05%)	398,827	410,833 (3.01%)	260,243	288,869 (11.00%)
MultiEMO	527,617	577,957 (9.54%)	378,345	390,351 (3.17%)	239,761	268,387 (11.94%)
RJCMA	557,601	607,941 (9.03%)	399,081	411,087 (3.01%)	260,497	289,123 (10.99%)
DMVAE	548,003	598,011 (9.13%)	394,603	406,607 (3.04%)	256,019	284,643 (11.18%)
TCGAN	522,086	572,096 (9.58%)	378,477	390,483 (3.17%)	239,893	268,519 (11.93%)
DCCA	521,921	571,931 (9.58%)	378,089	390,095 (3.18%)	239,505	268,131 (11.95%)
MVCN	537,857	588,197 (9.36%)	386,537	398,543 (3.11%)	247,953	276,579 (11.54%)
Dropout	529,077	579,117 (9.46%)	379,257	391,263 (3.17%)	240,673	269,299 (11.89%)
AWD	528,267	578,297 (9.47%)	378,737	390,743 (3.17%)	240,153	268,779 (11.92%)

HLGM	530,697	580,757 (9.43%)	380,297	392,303 (3.16%)	241,713	270,339 (11.84%)
SLMV	527,457	578,117 (9.60%)	378,217	390,223 (3.17%)	239,633	268,259 (11.95%)
MFIR	527,297	577,637 (9.55%)	382,313	394,319 (3.14%)	243,729	272,355 (11.75%)
MoEF	528,422	578,762 (9.53%)	378,861	390,867 (3.17%)	240,277	268,903 (11.91%)
MMoE	527,788	578,128 (9.54%)	378,482	390,488 (3.17%)	239,898	268,524 (11.93%)

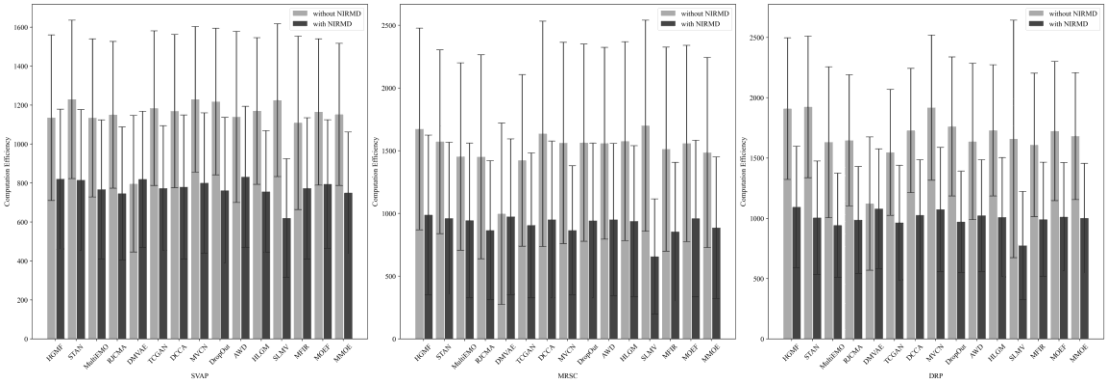

Figure G1. Time Complexity of Benchmarks with vs without NIRMD.

Appendix H. Experiment Results

Table H1. Mean (Standard Deviation) of RMSE with vs without NIRMD in the SVAP Case.

Multimodal Deep Learning Method	Feature Set						Fusion Best
	F-Q	F-T	F-G	F-A	F-V	F	
HGMF	37.67 (0.52)	36.88 (0.68)	38.51 (0.71)	37.25 (0.61)	38.51 (0.36)	39.83 (0.73)	×
HGMF+NIRMD	36.87 (0.77)	35.63 (0.54)	35.06 (0.81)	36.42 (0.76)	35.57 (0.69)	**34.80 (0.54)**	√
STAN	38.48 (0.28)	38.14 (0.69)	38.10 (0.33)	38.38 (0.72)	39.63 (0.33)	38.73 (0.72)	×
STAN+NIRMD	35.89 (0.45)	35.87 (0.87)	36.64 (0.49)	36.42 (0.86)	35.25 (0.84)	**34.89 (0.95)**	√
MultiEMO	31.82 (3.80)	31.12 (3.89)	31.16 (4.00)	32.02 (3.86)	32.67(4.49)	33.16 (4.27)	×
MultiEMO+NIRMD	25.13 (1.36)	25.81 (2.01)	25.54 (1.65)	25.95 (2.04)	26.09(2.50)	**23.60 (1.37)**	√
RJCMA	34.27 (8.65)	35.36 (8.81)	35.20 (9.38)	34.98 (8.98)	33.99 (8.07)	35.52 (8.70)	×
RJCMA+NIRMD	26.56 (0.78)	27.13 (0.68)	28.17 (0.79)	28.10 (1.13)	28.45 (0.97)	**25.56 (0.89)**	√
DMVAE	33.42 (0.42)	35.21 (0.72)	33.85 (0.45)	36.91 (0.88)	32.82 (0.71)	33.33 (0.83)	×
DMVAE+NIRMD	30.33 (0.79)	32.39 (0.80)	30.35 (0.83)	34.36 (0.83)	29.38 (0.78)	**28.34 (0.80)**	√
TCGAN	38.72 (0.34)	38.67 (0.31)	38.51 (0.32)	38.69 (0.32)	38.71 (0.32)	38.93 (0.33)	×
TCGAN+NIRMD	37.03 (0.54)	36.44 (0.45)	33.28 (0.52)	34.17 (0.72)	36.14 (0.67)	**33.10 (0.49)**	√
DCCA	32.19 (0.69)	35.78 (0.69)	34.08 (0.96)	36.98 (0.97)	31.94 (1.24)	33.28 (1.21)	×
DCCA+NIRMD	28.76 (0.98)	32.47 (0.82)	31.47 (0.76)	33.38 (0.77)	28.56 (1.08)	**28.40 (0.77)**	√
MVCN	35.08 (4.70)	35.31 (5.55)	36.53 (5.51)	34.53 (4.94)	34.75 (4.98)	36.13 (4.97)	×
MVCN+NIRMD	24.61 (0.70)	24.05 (0.64)	24.49 (0.73)	24.61 (0.66)	24.62 (0.72)	**22.55 (0.64)**	√
Dropout	33.73 (3.27)	33.54 (3.02)	33.69 (3.40)	33.28 (3.18)	33.54 (3.24)	33.39 (3.88)	×
Dropout+NIRMD	32.33 (1.10)	31.53 (0.96)	32.45 (0.92)	32.55 (0.96)	32.67 (0.86)	**31.52 (1.03)**	√
AWD	33.33 (4.54)	34.05 (4.89)	33.61 (4.33)	32.76 (4.56)	34.22 (4.42)	36.08 (5.15)	×
AWD+NIRMD	32.19 (1.01)	31.58 (1.19)	32.53 (0.98)	32.52 (1.17)	32.70 (1.24)	**31.54 (1.05)**	√
HLGM	34.13 (8.24)	35.33 (8.56)	35.15 (8.44)	35.99 (8.42)	34.81 (8.63)	35.41 (9.05)	×
HLGM+NIRMD	31.86 (0.84)	31.25 (0.87)	31.77 (0.87)	31.79 (0.90)	31.90 (0.87)	**30.97 (1.00)**	√
SLMV	31.53 (0.26)	31.53 (0.28)	31.54 (0.30)	31.56 (0.28)	31.56 (0.29)	31.53 (0.28)	√
SLMV+NIRMD	24.69 (0.65)	24.10 (0.63)	24.47 (0.71)	24.73 (0.75)	24.54 (0.74)	**22.60 (0.64)**	√
MFIR	29.26 (1.83)	29.03 (1.74)	29.28 (1.92)	29.44 (2.51)	29.69 (2.13)	29.52 (2.15)	×
MFIR+NIRMD	24.11 (0.62)	24.07 (0.63)	24.09 (0.60)	24.05 (0.63)	24.06 (0.61)	**22.51 (0.63)**	√
MoEF	30.32 (2.21)	30.82 (3.73)	30.33 (2.70)	30.48 (2.10)	30.88 (2.68)	32.68 (6.22)	×
MoEF+NIRMD	27.16 (2.03)	27.27 (3.29)	27.72 (2.19)	27.78 (2.14)	27.46 (2.42)	**25.30 (2.04)**	√

MMoE	28.61 (2.45)	29.54 (4.29)	29.02 (4.28)	29.88 (5.10)	27.82 (2.28)	28.56 (3.56)	×
MMoE+NIRMD	24.59 (0.79)	25.18 (2.42)	24.81 (2.00)	24.97 (1.89)	25.08 (1.93)	**22.99 (0.84)**	√

Fusion Best: √ (×) indicates that the performance on the full multimodal feature set (F) was (not) the overall best compared to that on the ablated multimodal feature sets.

Table H2. Mean (Standard Deviation) of AUC (%) with vs without NIRMD in the MRSC Case.

Multimodal Deep Learning Method	Feature Set					Fusion Best
	F-Q	F-T	F-G	F-V	F	
HGMF	85.46 (1.37)	64.62 (2.18)	87.21 (1.41)	86.18 (1.54)	86.42 (1.57)	×
HGMF+NIRMD	86.62 (1.27)	65.20 (1.21)	88.38 (0.97)	87.94 (1.18)	**89.26 (1.75)**	√
STAN	88.17 (0.96)	63.99 (1.52)	88.30 (1.15)	88.16 (0.97)	87.10 (1.06)	×
STAN+NIRMD	89.99 (1.69)	64.80 (1.47)	89.99 (1.39)	90.19 (1.43)	**90.25 (1.65)**	√
MultiEMO	85.61 (2.18)	56.70 (3.67)	86.59 (2.07)	85.36 (2.74)	85.15 (2.68)	×
MultiEMO+NIRMD	90.44 (2.03)	61.19 (3.58)	90.74 (1.99)	90.08 (2.22)	**92.41 (1.97)**	√
RJCMA	82.38 (4.78)	57.71 (4.05)	86.32 (2.08)	81.16 (5.01)	80.86 (4.18)	×
RJCMA+NIRMD	83.68 (1.79)	64.79 (4.03)	87.76 (0.94)	85.93 (3.70)	**88.33 (2.26)**	√
DMVAE	82.97 (1.30)	62.75 (2.03)	83.60 (1.60)	82.93 (1.37)	83.01 (1.87)	×
DMVAE+NIRMD	86.45 (1.35)	63.09 (1.74)	86.63 (1.67)	86.24 (1.16)	**87.11 (1.46)**	√
TCGAN	87.84 (1.00)	64.75 (1.69)	88.10 (0.94)	87.97 (1.08)	87.83 (1.16)	×
TCGAN+NIRMD	88.81 (0.87)	65.69 (0.92)	89.77 (0.69)	89.39 (0.95)	**89.96 (1.43)**	√
DCCA	86.22 (1.35)	64.15 (1.10)	87.47 (1.18)	86.51 (1.15)	86.49 (1.41)	×
DCCA+NIRMD	89.89 (1.14)	65.16 (1.78)	89.85 (1.42)	89.45 (1.86)	**90.09 (1.31)**	√
MVCN	87.23 (2.07)	58.83 (2.64)	87.65 (1.86)	87.36 (1.96)	87.39 (2.02)	×
MVCN+NIRMD	91.06 (2.22)	62.24 (2.74)	91.32 (1.86)	90.93 (1.92)	**92.98 (1.97)**	√
Dropout	86.95 (2.29)	61.96 (2.89)	87.76 (2.41)	87.13 (2.28)	87.73 (2.60)	×
Dropout+NIRMD	88.41 (1.28)	67.13 (4.33)	88.06 (0.41)	88.85 (0.98)	**89.40 (0.42)**	√
AWD	87.04 (2.53)	62.00 (2.89)	88.16 (2.49)	86.85 (2.35)	87.51 (3.06)	×
AWD+NIRMD	88.54 (0.89)	67.78 (1.59)	88.51 (1.03)	88.59 (1.66)	**89.28 (1.12)**	√
HLGM	85.53 (2.81)	57.60 (4.19)	86.35 (2.35)	85.87 (2.98)	85.99 (2.93)	×
HLGM+NIRMD	87.00 (1.33)	59.62 (4.94)	87.70 (1.68)	87.60 (1.83)	**88.06 (1.97)**	√
SLMV	87.38 (2.17)	58.74 (2.68)	87.95 (1.96)	87.60 (1.95)	87.46 (2.10)	×
SLMV+NIRMD	91.10 (2.00)	62.35 (2.65)	91.24 (1.86)	90.97 (1.88)	**92.82 (1.84)**	√
MFIR	87.63 (2.10)	58.79 (2.67)	87.77 (1.98)	87.34 (2.27)	87.42 (2.15)	×
MFIR+NIRMD	91.13 (1.90)	62.21 (2.68)	91.07 (1.96)	91.05 (1.86)	**92.80 (1.91)**	√
MoEF	72.31 (7.42)	53.91 (4.09)	73.17 (11.59)	69.99 (9.89)	72.00 (6.80)	×
MoEF+NIRMD	83.50 (4.27)	58.85 (4.08)	86.46 (3.75)	81.46 (4.65)	**87.96 (3.75)**	√
MMoE	84.32 (3.05)	56.80 (4.01)	86.15 (2.31)	83.32 (3.41)	83.19 (3.76)	×
MMoE+NIRMD	89.18 (2.72)	61.60 (3.55)	89.99 (2.38)	89.18 (2.21)	**91.52 (2.30)**	√

Fusion Best: √ (×) indicates that the performance on the full multimodal feature set (F) was (not) the overall best compared to that on the ablated multimodal feature sets.

Table H3. Mean (Standard Deviation) of AUC (%) with vs without NIRMD in the DRP Case.

Multimodal Deep Learning Method	Feature Set					Fusion Best
	F-H	F-S	F-T	F-G	F	
HGMF	72.97 (1.63)	78.06 (2.22)	83.16 (1.86)	83.24 (1.68)	83.12 (1.75)	×
HGMF+NIRMD	74.03 (1.64)	79.64 (2.05)	83.71 (1.87)	84.17 (2.47)	**85.75 (2.07)**	√
STAN	71.61 (1.15)	74.58 (2.31)	82.59 (3.45)	77.74 (2.26)	81.14 (2.44)	×
STAN+NIRMD	72.27 (1.65)	76.00 (2.11)	83.68 (1.84)	83.85 (1.97)	**84.07 (1.95)**	√
MultiEMO	74.56 (2.40)	69.09 (2.55)	77.42 (2.82)	74.50 (2.59)	73.99 (2.57)	×
MultiEMO+NIRMD	78.01 (2.43)	72.90 (2.77)	82.15 (2.70)	77.90 (2.51)	**83.65 (2.70)**	√
RJCMA	71.09 (2.30)	75.55 (2.61)	80.28 (2.82)	76.80 (2.82)	75.55 (2.90)	×
RJCMA+NIRMD	75.53 (2.60)	78.93 (2.99)	81.60 (4.45)	80.76 (2.93)	**83.93 (3.13)**	√

DMVAE	73.47 (1.97)	73.07 (1.92)	83.81 (2.14)	84.02 (1.96)	82.49 (1.84)	×
DMVAE+NIRMD	74.77 (1.44)	74.89 (2.38)	84.69 (1.95)	85.21 (1.42)	**85.68 (2.64)**	√
TCGAN	72.32 (1.01)	78.62 (2.88)	83.39 (2.20)	83.83 (2.10)	83.10 (2.65)	×
TCGAN+NIRMD	73.54 (1.30)	79.97 (2.66)	84.43 (1.38)	85.03 (2.14)	**85.63 (2.48)**	√
DCCA	72.42 (1.61)	76.57 (2.81)	80.52 (2.00)	81.90 (2.14)	81.82 (2.72)	×
DCCA+NIRMD	73.56 (1.89)	77.61 (2.60)	81.01 (1.57)	83.60 (2.17)	**84.36 (2.19)**	√
MVCN	68.89 (1.82)	68.15 (1.22)	70.08 (4.05)	68.57 (0.98)	69.06 (1.67)	×
MVCN+NIRMD	76.85 (2.22)	73.00 (1.92)	84.90 (2.08)	77.22 (2.65)	**86.40 (2.08)**	√
Dropout	72.09 (2.54)	76.59 (2.33)	80.22 (2.10)	78.16 (2.44)	79.53 (2.90)	×
Dropout+NIRMD	74.26 (2.19)	79.06 (2.56)	82.39 (2.62)	81.12 (2.67)	**84.63 (2.71)**	√
AWD	71.33 (2.34)	75.16 (2.31)	81.94 (2.11)	74.74 (2.65)	74.49 (2.48)	×
AWD+NIRMD	74.47 (3.14)	80.96 (2.50)	82.29 (1.90)	81.47 (3.10)	**83.41 (2.27)**	√
HLGM	73.76 (2.83)	77.97 (2.63)	81.07 (2.54)	77.50 (2.58)	76.48 (2.52)	×
HLGM+NIRMD	74.70 (2.32)	81.40 (2.77)	82.38 (2.85)	80.77 (2.80)	**83.26 (2.81)**	√
SLMV	70.34 (3.32)	68.66 (1.60)	71.58 (4.98)	71.24 (3.82)	69.45 (3.18)	×
SLMV+NIRMD	76.88 (2.56)	73.48 (2.10)	85.32 (1.90)	77.57 (2.96)	**86.82 (1.90)**	√
MFIR	73.38 (2.56)	69.52 (2.20)	81.80 (2.16)	73.69 (2.61)	72.94 (2.55)	×
MFIR+NIRMD	77.06 (2.71)	73.02 (2.32)	85.48 (1.98)	77.26 (2.56)	**86.98 (1.98)**	√
MoEF	72.32 (2.90)	69.57 (2.58)	74.45 (2.78)	74.31 (2.64)	72.96 (2.88)	×
MoEF+NIRMD	76.23 (2.93)	73.25 (2.80)	78.67 (2.70)	78.11 (2.28)	**80.31 (2.52)**	√
MMoE	73.67 (3.33)	69.17 (3.02)	78.68 (3.44)	74.62 (2.65)	74.16 (2.93)	×
MMoE+NIRMD	77.82 (3.48)	72.61 (2.74)	82.23 (3.47)	78.82 (3.11)	**83.78 (3.33)**	√

Fusion Best: √ (×) indicates that the performance on the full multimodal feature set (F) was (not) the overall best compared to that on the ablated multimodal feature sets.

References in the Online Supplement

Devlin J, Chang M-W, Lee K, Toutanova K (2018) BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:04805*.

Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet? *Proceedings of the IEEE Conference on CVPR*, 6546-6555.

Knight, K., & Fu, Y. (2000). On the efficiency of lasso-type estimators. *The Annals of Statistics* 28(2), 787-811.

Liu X, You X, Zhang X, Wu J, Lv P (2020) Tensor graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 8409-8416.

Petty BE, Gillespie AI, Shelly S, Klein AM (2020) Beauty and attractiveness in the human voice. *J. Voice* 36(4), 507-514.

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-CAM: Visual explanations from deep networks via gradient-based localization. *ICCV*, 618-626.

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, 5998-6008.

Yao L, Mao C, Luo Y (2019) Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 7370-7377.

Zhao, Y., & Yu, B. (2006). On model selection consistency of Lasso. *Journal of Machine Learning Research*, 11, 1205-1235.

Zheng F, Zhang G, Song Z (2001) Comparison of different implementations of MFCC. *Journal of Computer Science and Technology* 16(6):582-589.