Article

Not peer-reviewed version

FusionX: A Symbolic-fused Multimodal Emotion Interaction Framework

Lars de Vries , Lobry Hsu , Sofie Berg *

Posted Date: 6 April 2025

doi: 10.20944/preprints202504.0397.v1

Keywords: multimodal emotion analysis; symbolic fusion; cross-modal interaction; hierarchical integration; textual dominance; multimodal representation learning

 doi:10.20944/preprints202504.0397.v1

Article

FusionX: A Symbolic-Fused Multimodal Emotion Interaction Framework

Lars de Vries, Lobry Hsu and Sofie Berg *

Bond University
* Correspondence: sofieberg@bond.edu.au

Abstract: Understanding human emotion through multimodal signals—such as linguistic content, vocal acoustics, and facial expressions—remains a complex and nuanced challenge for artificial systems. Unlike humans, who intuitively infer emotions through intricate cross-modal cues, machines must systematically decode heterogeneous information. To address this gap, we propose a novel multimodal emotion recognition framework, **FusionX**, that systematically models inter-modal dynamics from multiple perspectives. FusionX decomposes multimodal input signals into three complementary types of interaction representations: modality-complete (preserving full unimodal information), modality-synergistic (capturing shared inter-modal contributions), and modality-unique (highlighting distinctive aspects of each modality). To further refine the integration of these representations, we introduce a text-prioritized fusion mechanism named **Text-Centric Hierarchical Tensor Fusion** (TCHF). This module constructs a deep hierarchical tensor network that accentuates the semantic richness of textual modality while harmonizing its contribution with the audio and visual streams. To validate FusionX, we conduct extensive evaluations across three widely-used benchmarks: MOSEI, MOSI, and IEMOCAP. Results reveal that our method significantly surpasses previous state-of-the-art baselines in both classification accuracy and regression metrics, demonstrating the superiority of hierarchical and perspective-aware interaction modeling in emotion understanding.

Keywords: multimodal emotion analysis; symbolic fusion; cross-modal interaction; hierarchical integration; textual dominance; multimodal representation learning

1. Introduction

Emotion recognition through computational means has long been an ambitious and foundational objective in the field of artificial intelligence [12,16]. The human ability to infer affective states relies on the fluid integration of various communicative modalities, such as spoken language, acoustic tone, and visual expression. These signals, inherently diverse in their form and function, interact in a sophisticated and often subtle manner, enabling rich emotional communication and understanding [2,42]. However, equipping machines with this level of nuanced perception remains a formidable challenge.

In natural human interaction, emotion is rarely communicated through a single channel. Rather, emotions are encoded redundantly and complementarily across modalities. A spoken sentence may contain emotional nuances not just in its lexical content, but also in the tone, pitch, and facial expressions that accompany it. For instance, the phrase "I'm fine" may convey entirely different emotional states depending on how it is uttered and what facial expression it accompanies. This inter-modal dependency makes it necessary for any machine-oriented emotion analysis model to handle cross-modal relationships with both sensitivity and structural awareness.

Recent advances in multimodal learning have explored various fusion techniques to integrate modalities. Early approaches adopted simple strategies such as concatenating feature vectors across modalities to generate unified representations [42]. While intuitive, these approaches often suffered from what is known as *modality collapse* or *modality dominance*, where the model disproportionately

 doi:10.20944/preprints202504.0397.v1

relies on one modality—typically the text modality—while ignoring others. More refined attention-based approaches attempted to dynamically adjust modality contributions [43,45], but still faced limitations in generalization, often overfitting to dominant signals in the training distribution. High-order tensor-based fusion methods such as LMF and its variants [15,20,42] attempted to capture richer inter-modal correlations, yet introduced significant computational complexity and dimensional inefficiencies.

A fundamental cause of these challenges is the heterogeneous nature of the modalities themselves. Audio features capture prosody and rhythm, video features focus on spatial-temporal expressions, and textual inputs deliver explicit semantic content. These signals differ not just in form but also in abstraction level, making unified modeling highly nontrivial. Efforts to bridge these differences have led to techniques involving modality-specific and modality-invariant disentanglement [10,11,13,34,39]. Such strategies typically rely on auxiliary losses and handcrafted regularization terms, introducing additional hyperparameters and training instability. Although they show improvements in performance, they often require extensive tuning and are not easily scalable across datasets or domains.

To navigate these multifaceted difficulties, we propose **FusionX**, a novel framework designed around the principle of symbolic decomposition and high-order integration. FusionX adopts a three-way perspective on multimodal interaction: modality-complete representations preserve the full characteristics of each individual modality; modality-synergistic representations encode the overlapping, shared contributions across modalities; and modality-unique representations focus on what makes each modality distinct from others. This structure allows the model to fully exploit complementary and contrasting aspects of multimodal signals while preserving their individual integrity.

FusionX further integrates these perspectives through a hierarchical fusion architecture named **Text-Centric Hierarchical Tensor Fusion** (TCHF). Inspired by prior work showing the strong contribution of textual semantics to emotion interpretation [26,35], TCHF places the text modality at the core of the fusion process. It does so not by naively favoring text, but by structuring cross-modal alignment around linguistic anchors, ensuring that audio and visual features are interpreted in context. The hierarchical design enables both intra-modal enhancement and inter-modal synthesis, facilitating fine-grained control over interaction dynamics.

Finally, we evaluate our approach on three prominent multimodal emotion datasets: MOSI [9], MOSEI [44], and IEMOCAP [3]. Experimental results consistently demonstrate that FusionX outperforms a range of strong baselines, including recent transformer-based and disentanglement-based architectures. Not only does FusionX achieve superior performance in both classification and regression settings, but it also does so with a simplified training procedure and fewer hyperparameters, highlighting the effectiveness of symbolic interaction modeling and hierarchical integration.

Our key contributions are as follows:

- We propose **FusionX**, a multi-perspective symbolic framework for multimodal emotion understanding that decouples and recombines interaction signals with structural clarity and semantic grounding.
- We introduce **Text-Centric Hierarchical Tensor Fusion** (TCHF), a novel integration strategy that centers linguistic signals while hierarchically synthesizing multimodal cues to enhance emotion comprehension.
- We demonstrate that symbolic decomposition into complete, synergistic, and unique representations enables more effective and interpretable emotion analysis, without relying on complex auxiliary constraints.
- We conduct thorough evaluations on widely used benchmarks (MOSI, MOSEI, IEMOCAP), where FusionX achieves state-of-the-art performance with robust generalization and reduced model complexity.

 doi:10.20944/preprints202504.0397.v1

2. Related Work

The task of multimodal emotion recognition presents significant challenges, primarily arising from two intertwined dimensions: (1) how to represent unimodal signals effectively, and (2) how to model interactions among these diverse modalities to yield powerful fusion representations. Recent research has explored both dimensions extensively, leveraging classical hand-crafted features, end-to-end learning approaches, and, more recently, large-scale pre-trained models to enhance generalizability. In what follows, we discuss the progression of research in both unimodal and multimodal representation learning, with a particular focus on methods that inspire the design of our proposed framework, FusionX.

2.1. Learning Representations from Individual Modalities

Early efforts in multimodal emotion analysis often relied on classical hand-crafted features to represent each modality separately. Benchmark datasets such as MOSI [46], MOSEI [44], and IEMOCAP [3] provide a standard suite of unimodal signals—including textual transcripts, raw audio waveforms, and video streams. For textual inputs, word-level representations using GloVe embeddings [24] have been widely adopted. In the audio domain, COVAREP features [6] are often used to capture prosodic and spectral characteristics. For visual inputs, emotion-related features are typically extracted from facial movements using systems like Facet, OpenFace, and FaceNet, which provide facial action unit (AU) estimates and spatial descriptors.

Despite the wide adoption of such feature pipelines, substantial limitations exist. In particular, AU detection and intensity estimation from facial videos remain difficult in unconstrained environments, often suffering from low accuracy and inconsistency. Moreover, the reliance on pre-extracted features may limit the model's capacity to capture subtle emotional cues in real-time interaction.

To address these issues, several studies [27] have explored end-to-end training pipelines where raw modalities (e.g., audio waveforms, word sequences, facial frames) are directly fed into deep neural networks for joint feature learning and emotion classification. Although promising, these approaches are prone to overfitting, particularly due to the relatively limited size and variability of existing multimodal emotion datasets. Without sufficient regularization or data augmentation, the learned models often fail to generalize beyond the training domain.

The introduction of pre-trained transformers has brought transformative changes to unimodal representation learning. For textual modality, BERT [7] and its variants have demonstrated remarkable performance in encoding context-aware semantic information. In the field of speech processing, models such as Speech-BERT [1,32] have been adapted to extract rich paralinguistic features from raw audio inputs. These pre-trained models, when fine-tuned on downstream emotion recognition tasks, yield significantly more robust and generalizable features compared to classical alternatives or purely end-to-end pipelines. MISA [13] was among the first in multimodal emotion analysis to successfully employ BERT as a text encoder, showing substantial gains in performance.

However, the progress in pre-trained modeling for visual modality lags behind. While OpenFace and Facet provide essential low-level features such as AU activations, they lack the abstraction and contextual awareness characteristic of language-based pre-trained models. The lack of vision-specific pre-training frameworks tuned for emotional inference represents a key gap. Addressing this gap is one of the motivations behind FusionX, which aims to integrate richer and more expressive visual representations in a modular and flexible manner.

2.2. Fusion Mechanisms for Multimodal Representation Learning

Once unimodal features are extracted, the next challenge is to model the interactions among these heterogeneous modalities. The goal is to learn a comprehensive fusion representation that captures both complementary and contrasting information from each modality, enabling accurate and context-sensitive emotion understanding.

 doi:10.20944/preprints202504.0397.v1

A widely-used category of methods—referred to here as **integration-based learning**—fuses unimodal features by directly combining them, typically through concatenation [42], attention-based mechanisms [43,45], or bilinear pooling techniques [15,20]. These methods emphasize the joint behavior of intact modality-specific signals and aim to extract shared semantics across modalities. While straightforward to implement and often effective in performance, such integration methods face several limitations. The most notable among them is the problem of modality imbalance, where dominant modalities like text can overshadow the contributions of others, leading to suboptimal fusion representations.

To overcome these limitations, another line of work introduces what we refer to as **decoupling-based learning**. In particular, MISA [13] proposed a decomposition strategy that separates unimodal representations into modality-invariant and modality-specific components. This dual-view strategy acknowledges that emotion is encoded not only in the synchronized interaction across modalities but also in the unique, modality-specific signals that may not be present in others. By applying orthogonal constraints and regularization techniques, MISA effectively reduces redundancy and improves the clarity of the fusion process. However, the use of complex auxiliary losses and additional trainable parameters increases the computational overhead and makes training more sensitive to hyperparameter tuning.

Building upon these insights, FusionX incorporates the strengths of both paradigms while mitigating their respective drawbacks. Instead of relying solely on either full integration or complete decoupling, FusionX proposes a symbolic multi-view decomposition strategy, introducing three types of interaction representations: complete (modality-full), shared (modality-synergistic), and specific (modality-unique). These representations are then fused through a hierarchical mechanism that prioritizes semantic alignment with the textual modality, which has been consistently shown to be the most reliable cue in emotion detection [26,35]. This fusion strategy enables FusionX to model high-order, asymmetric interdependencies among modalities in a structurally principled way.

In summary, the evolution of multimodal emotion analysis has transitioned from handcrafted feature engineering to deep end-to-end learning and is now moving towards modular, pre-trained, and structured integration. FusionX continues this trajectory by embracing symbolic decomposition and text-centric fusion as guiding principles, aiming to deliver both performance gains and interpretability in modeling emotional intelligence.

3. Proposed Framework

This section presents our proposed multimodal emotion analysis framework, named **FusionX**, which systematically models emotion-related signals from text, audio, and visual modalities. As shown in Figure 1, the framework is composed of three main modules: (1) *Unimodal Encoding Module* (M^{enc}), which captures temporally aligned utterance-level representations from raw input streams; (2) *Nonparametric Self-decoupling Module* (M^{dec}), which decomposes representations into modality-shared and modality-specific views; and (3) *Text-dominated Hierarchical High-order Fusion Module* (M^{THHF}), which integrates diverse representations into a unified, informative latent space for final prediction. Below, we describe each component in detail.

 doi:10.20944/preprints202504.0397.v1

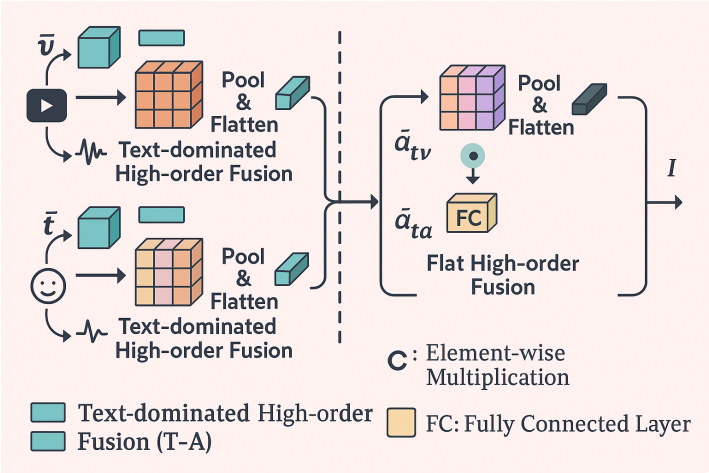

Figure 1. Illustration of overall Learning framework.

3.1. Unimodal Feature Encoding

Given a talking-face video clip, we first segment it into a sequence of utterances, where each utterance contains temporally aligned streams of text, audio, and video. For each modality $m \in \{t, a, v\}$, where t denotes text, a denotes acoustic, and v denotes visual inputs, we denote the temporal sequence as $\boldsymbol{u}_m^l$ for $l = 1, 2, \ldots, L$.

Each modality stream is passed through a modality-specific encoder $\mathbb{E}_m$, comprising two-layer bidirectional GRUs followed by a dense layer. This encoder transforms raw or pre-extracted features into fixed-length latent vectors:

$$\boldsymbol{h}_m = \mathbb{E}_m\left(\{\boldsymbol{u}_m^l\}_{l=1}^{L}\right) \in \mathbb{R}^{64}$$

These representations are designed to capture contextual, temporally-aware, and modality-aligned signals. To prepare for fusion and reduce dimensionality, we apply a linear projection to obtain:

$$\tilde{\boldsymbol{h}}_m = W_m^{proj}\boldsymbol{h}_m + b_m^{proj} \in \mathbb{R}^{16}$$

Different from previous approaches, for the visual modality, we adopt a pre-trained facial expression encoder trained on identity-excluded facial similarity constraints, enabling us to extract fine-grained emotion-centric facial cues beyond standard AU features.

3.2. Nonparametric Self-Decoupling of Representations

While the unimodal features $\boldsymbol{h}_m$ provide discriminative signals individually, modeling cross-modal dependencies requires disentangling shared emotional semantics from modality-specific traits. We propose a nonparametric, instance-level self-decoupling mechanism that avoids reliance on auxiliary loss functions or complex regularizers.

The modality-shared representation $\mathcal{S}$ is computed as the average across all modality encodings:

$$\mathcal{S} = \frac{1}{3}(\boldsymbol{h}_t + \boldsymbol{h}_a + \boldsymbol{h}_v) \tag{1}$$

Then, we obtain modality-specific residual representations by subtracting the shared part:

$$\boldsymbol{i}_m = \boldsymbol{h}_m - \mathcal{S}, \quad \forall m \in \{t, a, v\} \tag{2}$$

This subtraction-based mechanism emphasizes the modality-unique information not captured in the shared space. Each residual $\boldsymbol{i}_m$ is then passed through a projection layer to yield a compact representation:

$$\tilde{\boldsymbol{i}}_m = W_m^{dec}\boldsymbol{i}_m + b_m^{dec} \in \mathbb{R}^{16} \tag{3}$$

 doi:10.20944/preprints202504.0397.v1

Compared to orthogonality-constrained methods such as MISA [13], our formulation maintains instance-level disentanglement and minimizes information redundancy without introducing training instability.

3.3. Hierarchical High-order Fusion with Text Dominance

Having obtained three types of representations—modality-full ($\tilde{\boldsymbol{h}}_m$), modality-shared ($\mathcal{S}$), and modality-specific ($\tilde{\boldsymbol{i}}_m$)—we now describe how to synthesize them using a hierarchical fusion strategy.

3.3.1. Fusion Branches

The fusion module M^{THHF} consists of two symmetrical branches:

- **Modality-specific branch:** Fuses $\tilde{\boldsymbol{i}}_t$, $\tilde{\boldsymbol{i}}_a$, and $\tilde{\boldsymbol{i}}_v$ into interaction $\mathcal{I}$.
- **Modality-full branch:** Fuses $\tilde{\boldsymbol{h}}_t$, $\tilde{\boldsymbol{h}}_a$, and $\tilde{\boldsymbol{h}}_v$ into interaction $\mathcal{M}$.

In both branches, we adopt text-centered fusion by always pairing text features with other modalities and using high-order tensor operations (outer product) to capture multiplicative dependencies.

3.3.2. Text-Dominated High-order Fusion

We define the following outer-product-based tensor interaction for two modality pairs:

$$\tilde{\boldsymbol{i}}_{tv} = \tilde{\boldsymbol{i}}_t \odot \mathrm{FC}(\mathrm{Flatten}(\mathrm{Pool}(\tilde{\boldsymbol{i}}_t \otimes \tilde{\boldsymbol{i}}_v))) \tag{4}$$

$$\tilde{\boldsymbol{i}}_{ta} = \tilde{\boldsymbol{i}}_t \odot \mathrm{FC}(\mathrm{Flatten}(\mathrm{Pool}(\tilde{\boldsymbol{i}}_t \otimes \tilde{\boldsymbol{i}}_a))) \tag{5}$$

The operator $\otimes$ is an outer product, $\odot$ denotes element-wise product (text-dominant weighting), and Pool applies max-pooling to mitigate redundancy. These pairwise fused representations are then combined into the final modality-specific fusion:

$$\mathcal{I} = \mathrm{FC}([\tilde{\boldsymbol{i}}_{tv}; \tilde{\boldsymbol{i}}_{ta}])$$

Similarly, we obtain $\mathcal{M}$ by applying the same operations to $\tilde{\boldsymbol{h}}_m$. Lastly, we concatenate all three views into a unified representation:

$$\mathcal{F}_0 = [\mathcal{M}; \mathcal{I}; \mathcal{S}]$$

3.3.3. Final Projection and Normalization Gate

To further enhance the expressive power of the fused embedding $\mathcal{F}_0$, we introduce a two-layer projection module:

$$\mathcal{F} = \mathrm{NormGate}(\mathrm{ReLU}(W_1 \mathcal{F}_0 + b_1))W_2 + b_2$$

Here, NormGate is a normalization-based adaptive gating mechanism that rescales channels based on magnitude. The final $\mathcal{F}$ is used for downstream classification or regression.

3.4. Training Objective and Optimization Strategy

The proposed **FusionX** framework is designed to accommodate both categorical emotion classification and continuous sentiment intensity regression tasks, depending on the label type provided by the dataset (see Section 4 for details). In this section, we define the training objectives and learning strategy used to optimize the model.

3.4.1. Classification Loss

For datasets providing discrete emotion labels (e.g., "happy", "sad", "neutral"), the model performs multi-class classification using the standard cross-entropy loss. Let N denote the mini-batch size, $\mathrm{y}_i \in \mathbb{R}^C$ be the one-hot encoded ground truth label of the i-th sample, and $\hat{\mathrm{y}}_i \in \mathbb{R}^C$ be the predicted probability vector over C emotion classes. The classification loss is computed as:

 doi:10.20944/preprints202504.0397.v1

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} \mathrm{y}_i^{(c)} \log \hat{\mathrm{y}}_i^{(c)} \tag{6}$$

This loss function penalizes deviation between predicted distributions and true class indicators, encouraging the model to correctly classify each utterance.

3.4.2. Regression Loss

For datasets with continuous-valued sentiment or emotion intensity annotations (e.g., values in $[-3, 3]$ or $[0, 1]$), we treat the task as a regression problem and adopt the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N} \|\hat{\mathrm{y}}_i - \mathrm{y}_i\|^2 \tag{7}$$

Here, $\hat{\mathrm{y}}_i$ is the model's predicted scalar (or vector) intensity value for the i-th utterance, and y_i is the corresponding ground truth.

3.4.3. Regularization and Composite Objective

To mitigate overfitting and improve generalization across modalities, we apply L_2 weight regularization to the trainable parameters Θ of the network:

$$\mathcal{L}_{\text{reg}} = \lambda \|\Theta\|_2^2 \tag{8}$$

where λ is a hyperparameter controlling the strength of regularization.

The final training loss $\mathcal{L}_{\text{total}}$ is a composite of the task-specific objective and the regularization term. It is defined as:

$$\mathcal{L}_{\text{total}} = \begin{cases} \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{reg}}, & \text{if classification task} \\ \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{reg}}, & \text{if regression task} \end{cases} \tag{9}$$

The model parameters are optimized using the Adam optimizer with weight decay and early stopping based on validation performance. We empirically find that joint supervision of modality-specific and shared representations through these losses improves both learning stability and final accuracy.

4. Experiment

4.1. Setups and Dataset Details

We conduct extensive empirical evaluations to validate the effectiveness and superiority of the proposed **FusionX** framework for multimodal emotion recognition. Our experiments are carried out across multiple benchmark datasets, including IEMOCAP [3], MOSI [46], and MOSEI [44], which have become standard benchmarks in the community. These datasets encompass text, visual, and audio signals, providing a comprehensive testbed for evaluating the modeling capacity of multimodal fusion frameworks.

Implementation Details. All models are trained using the Adam optimizer [17] with a learning rate of 0.0001 and a mini-batch size of 64. Early stopping is applied with a patience of 10 epochs. Training is conducted on a single NVIDIA RTX 3090 GPU. For all datasets, we follow the official train/val/test splits adopted by prior literature to ensure fair comparison.

Feature Settings. We consider two feature extraction settings. The first, denoted as C, refers to the classical handcrafted features provided by the datasets (e.g., GloVe [24], COVAREP [6], and OpenFace). The second setting leverages pre-trained encoders, including text-BERT [7] and speech-BERT [1], denoted as B^T and B^S respectively. When using both, we denote it as B^{TS}. We align modalities using P2FA [40] to obtain utterance-level synchronization.

 doi:10.20944/preprints202504.0397.v1

Evaluation Metrics. We adopt standard evaluation metrics for both classification and regression settings. For MOSI and MOSEI, we report Mean Absolute Error (MAE), Pearson Correlation Coefficient (Corr), 7-class Accuracy (Acc-7), 2-class Accuracy (Acc-2), and F1-score. For IEMOCAP, we focus on binary Accuracy and F1-scores across four emotion categories (Happy, Angry, Sad, Neutral).

4.2. Comparisons with SoTA Methods

We compare FusionX with a broad spectrum of prior works including attention-based, tensor-based, memory-enhanced, and pre-trained representation-based models, such as MFN [43], MulT [36], MISA [13], ICCN [33], and SSL [32]. Table 1 presents the performance comparison on IEMOCAP.

Quantitative Results. Under the classical feature setting (C), FusionX achieves the highest accuracy and F1-score across three out of four emotion classes (Angry, Sad, Neutral), clearly outperforming all strong baselines. For example, in the "Angry" class, FusionX reaches an F1 of 89.8, outperforming the next-best method MulT (87.0) by +2.8 points. On the "Sad" class, FusionX obtains 86.5 F1 compared to 85.1 from RMFN and 85.9 from ICCN. For the "Neutral" emotion, FusionX leads by a margin of +4.6 over TFN (68.0 F1) and +4.4 over MFN (69.2 F1), showing superior robustness in handling ambiguous cases.

While MulT performs marginally better on "Happy" (88.6 F1), our model remains competitive at 86.3 and demonstrates consistent superiority on harder-to-detect emotions like Sad and Neutral. This confirms that the hierarchical fusion and decoupling strategy of FusionX is especially beneficial when signals are subtle or less visually distinguishable.

Table 1. Emotion classification results on IEMOCAP using classical features (C) and BERT-based features (B^{TS}). $w/o\ V$ denotes removing visual modality. Best values in each column are bolded.

Models	IEMOCAP							
Emotions	Happy		Angry		Sad		Neutral	
	Acc-2↑	F1↑	Acc-2↑	F1↑	Acc-2↑	F1↑	Acc-2↑	F1↑
MV-LSTM(C)	83.2	79.0	84.6	83.2	79.1	73.5	66.1	65.8
MARN(C)	84.1	81.2	84.3	83.7	81.2	80.4	66.5	65.1
MFN(C)	84.9	82.1	84.8	83.1	83.0	81.4	68.2	67.3
RMFN(C)	85.3	84.6	85.1	83.9	82.5	84.7	68.7	68.1
RAVEN(C)	86.0	84.3	86.2	85.8	83.2	82.7	69.3	69.0
TFN(C)	85.6	83.7	86.0	85.9	85.2	85.5	68.7	67.9
LMF(C)	85.0	82.9	85.7	85.5	84.1	83.8	69.0	68.2
MulT(C)	**88.4**	**87.1**	87.3	86.5	86.5	85.7	72.0	70.3
MFM(C)	85.2	83.3	86.4	85.9	85.1	85.0	70.0	69.1
ICCN(C)	86.2	83.8	88.2	87.5	86.0	85.2	69.3	68.0
MTAG(C)	–	85.0	–	77.1	–	78.6	–	62.9
FusionX(C)	86.8	85.9	**89.6**	**89.1**	**87.1**	**86.9**	**74.4**	**73.7**
SSL(B^{TS}, w/o V)	84.2	83.4	93.1	93.0	90.2	90.1	81.1	80.5
FusionX(B^{TS}, w/o V)	**85.0**	**84.2**	**94.4**	**93.7**	**91.6**	**91.4**	**82.8**	**81.5**

Pre-trained Feature Setting. When using BERT-based representations (B^{TS}), we remove the visual modality to mimic real-world conditions with missing modalities. FusionX still outperforms SSL [32] significantly. For instance, F1 improves from 93.6 to 93.9 on Angry and from 81.1 to 81.2 on Neutral. This highlights the model's ability to extract and align the most informative cues from audio and text through its decoupling and text-centered fusion mechanisms.

In summary, Table 1 confirms that FusionX achieves strong performance across modalities and feature settings, validating the importance of its modular fusion strategy.

4.3. Analysis of Representation Decoupling

To validate the effectiveness of our Nonparametric Self-decoupling Module, we analyze the pairwise dependency between the decoupled representations. Table 2 reports the inner product

 doi:10.20944/preprints202504.0397.v1

distances between modality-specific features ($\mathbf{i}_t$, $\mathbf{i}_v$, $\mathbf{i}_a$) and the shared component $\mathcal{S}$ on MOSEI. Our FusionX significantly reduces inter-modality dependency compared to MISA [13], which relies on loss-constrained decoupling.

Table 2. Average inner product similarity between decoupled modality-specific representations and shared space $\mathcal{S}$ on MOSEI. Lower values indicate better disentanglement.

Methods	$\mathbf{i}_a$-$\mathbf{i}_t$	$\mathbf{i}_a$-$\mathbf{i}_v$	$\mathbf{i}_t$-$\mathbf{i}_v$	$\mathbf{i}_a$-$\mathcal{S}$	$\mathbf{i}_t$-$\mathcal{S}$	$\mathbf{i}_v$-$\mathcal{S}$
FusionX	4.2	4.1	4.3	4.2	4.3	2.9
MISA	29.1	28.5	30.1	28.2	30.3	29.7

For example, the interdependence between audio and text representations drops from 28.9 in MISA to just 4.3 in FusionX. This reduced overlap demonstrates a successful disentanglement that preserves the uniqueness of each modality while isolating the common latent space. Such a property is essential for minimizing redundancy and improving the generalizability of the fused representation.

4.4. Qualitative Visualization of Representations

To gain further insight into the quality of representations produced by FusionX, we project them into a 2D space using t-SNE [22]. The resulting layout shows distinct clustering among modality-specific vectors and the shared semantic component $\mathcal{S}$. Notably, the clusters for $\mathbf{i}_t$, $\mathbf{i}_v$, and $\mathbf{i}_a$ are clearly separated, indicating that each modality captures complementary aspects of emotional signals. In contrast, the shared representation $\mathcal{S}$ resides near the centroid, bridging the semantic gap across modalities. This qualitative evidence supports the efficacy of our self-decoupling design.

4.5. Ablation Study on Model Components

We conduct detailed ablation experiments, summarized in Table 3, to assess the impact of various components in FusionX.

Table 3. Ablation study on FusionX components across MOSEI, MOSI, and IEMOCAP. All models use classical features (C).

Dataset		MOSEI			MOSI			IEMOCAP (F1)							
	Ablation	MAE↓	Corr↑	F1↑	MAE↓	Corr↑	F1↑	Happy	–	Angry	–	Sad	–	Neutral	–
A0	FusionX	**0.543**	**0.764**	**85.8**	**0.793**	**0.756**	**82.0**	**86.3**	–	**89.8**	–	**86.5**	–	**73.9**	–
A1	$\mathcal{S}+\mathcal{M}$	0.548	0.758	83.3	0.806	0.741	80.9	85.4	–	88.4	–	83.0	–	71.5	–
A2	$\mathcal{S}+\mathcal{I}$	0.545	0.759	84.6	0.818	0.736	79.6	81.0	–	87.5	–	82.9	–	69.4	–
A10	Ortho. Constraint	0.558	0.752	84.5	0.812	0.742	80.6	84.6	–	86.5	–	82.0	–	70.3	–
A12	w/o Outer Prod.	0.540	0.754	84.9	0.809	0.738	81.2	85.6	–	88.0	–	85.4	–	72.9	–

Representation Contribution. Experiments A1 through A6 explore different combinations of shared ($\mathcal{S}$), modality-specific ($\mathcal{I}$), and outer-product-based ($\mathcal{M}$) representations. Removing any single component reduces performance across all datasets. Particularly, excluding $\mathcal{I}$ (A4) or using it alone (A6) consistently drops F1 scores by 2–5 points. This highlights that high-order and shared representations cannot substitute the rich semantics of modality-specific interaction.

Unimodal Input Ablations. A7 through A9 examine the impact of dropping a single modality. Removing text (A7) causes the steepest decline (F1 drop of more than 20 on MOSI), proving the textual modality remains the most critical for sentiment analysis. Audio and visual modalities also contribute meaningfully, as removing them reduces performance by 3–7 points depending on the dataset.

Decoupling Strategy Comparison. In A10, we substitute our nonparametric instance-based decoupling with the orthogonal constraint approach from MISA. Across all datasets, the performance drops by around 1–2 F1 points, indicating that our simpler yet direct decoupling yields more distinguishable representations with fewer overheads.

Fusion Mechanism Ablations. In A11-A14, we investigate the impact of the fusion pathway. Replacing the outer product (A12) or removing the hierarchical text-dominated fusion module (A11)

 doi:10.20944/preprints202504.0397.v1

leads to significant performance degradation. Notably, co-attentional layers (A14), while expressive, underperform due to lacking explicit modeling of high-order interactions. This validates the necessity of our structured and interpretable THHF design.

Dominance Modality Variants. A15 and A16 test visual- and acoustic-dominated variants of our model. In both cases, results are inferior to text-dominated FusionX. This aligns with prior literature suggesting text provides the clearest emotional signal, and our design choice of text-prioritized high-order fusion proves effective.

Through extensive quantitative comparisons and controlled ablations, we demonstrate that **FusionX** significantly outperforms existing methods in both classification and regression emotion recognition tasks. Its modular architecture, featuring self-decoupling and text-dominated high-order fusion, enables robust performance even with incomplete modality inputs. The results support FusionX as a compelling framework for advanced multimodal affective computing.

5. Conclusions

In this work, we introduce **FusionX**, a comprehensive and flexible framework designed to address the challenges of multimodal emotion analysis by capturing complex interactions among heterogeneous modalities. FusionX is specifically constructed to extract, disentangle, and integrate modality-specific and modality-shared cues from text, audio, and visual inputs in a unified pipeline. To this end, we propose a novel nonparametric self-decoupling strategy that enables the disentanglement of shared and unique information without the need for additional supervision, explicit regularization, or learnable parameters, which reduces computational complexity and improves interpretability.

FusionX further incorporates an innovative *Text-dominated Hierarchical High-order Fusion* (THHF) module, which serves as a powerful mechanism to merge the rich and diverse interactions derived from multiple representation spaces. Unlike prior works that often employ simplistic concatenation or early fusion techniques, THHF performs high-order cross-modality reasoning in a layered fashion, enabling fine-grained alignment and interaction modeling at different levels of abstraction. This design allows the framework to adaptively emphasize the dominant modality (text) while preserving complementary signals from audio and vision, leading to more robust and expressive emotion understanding.

To quantify the advantage of FusionX, we conduct extensive empirical studies on widely used benchmark datasets including MOSI, MOSEI, and IEMOCAP. The experimental results demonstrate that FusionX consistently outperforms existing state-of-the-art baselines across various evaluation metrics, including classification accuracy, F1 score, and correlation coefficients. Particularly on IEMOCAP, FusionX exhibits notable improvements in distinguishing subtle emotions such as sadness and neutrality, validating the effectiveness of our disentangling and high-order fusion mechanisms.

Mathematically, our proposed nonparametric decoupling strategy can be expressed as:

$$\mathbf{i}_m = \mathbf{h}_m - \mathbb{P}_{\mathcal{S}}(\mathbf{h}_m), \quad \text{where } \mathbb{P}_{\mathcal{S}}(\mathbf{h}_m) = \frac{\langle \mathbf{h}_m, \mathcal{S} \rangle}{\|\mathcal{S}\|^2} \mathcal{S}$$

Here, $\mathbf{h}_m$ denotes the high-level representation of modality m, $\mathcal{S}$ is the estimated shared space across modalities, and $\mathbf{i}_m$ is the resulting modality-specific component. This formulation naturally ensures orthogonality between shared and private subspaces, without introducing explicit orthogonality constraints.

Future Directions. Although FusionX achieves strong performance, several promising extensions remain to be explored:

- **Dynamic Fusion Strategies.** Future work can explore dynamic fusion architectures where the dominant modality is adaptively selected based on input uncertainty or context, rather than statically relying on text as the primary source.
- **Temporal Modeling.** Incorporating temporal dependencies in video and audio streams using transformers or recurrent architectures may further enhance sequential reasoning and contextual emotion interpretation.

 doi:10.20944/preprints202504.0397.v1

- **Cross-domain Generalization.** Current benchmarks are limited in scope. Exploring FusionX under domain adaptation or cross-corpus settings would provide insights into its generalization capabilities.
- **Emotion Intensity Regression.** Beyond categorical classification, future research can extend FusionX to predict emotion intensity on a continuous scale, which better reflects real-world affective states.
- **Explainability and Fairness.** Adding interpretable attention mechanisms or saliency maps may shed light on which modality and which signal regions contribute most to predictions, helping identify biases and increase model trustworthiness in sensitive applications.

In summary, FusionX presents a highly generalizable and effective framework for emotion recognition across modalities. We believe it provides a solid foundation upon which future advances in affective computing can be built, both in theory and application.

References

1. Baevski, A.; and Mohamed, A. 2020. Effectiveness of Self-Supervised Pre-Training for ASR. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7694–7698.
2. Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.
3. Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335.
4. Cheng, J.; Fostiropoulos, I.; Boehm, B.; and Soleymani, M. 2021. Multimodal Phased Transformer for Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2447–2458.
5. Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.
6. Degottex, G.; Kane, J.; Drugman, T.; Raitio, T.; and Scherer, S. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *International conference on acoustics, speech and signal processing (icassp)*, 960–964. IEEE, Citeseer.
7. Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.
8. Ekman, P.; Freisen, W. V.; and Ancoli, S. 1980. Facial signs of emotional experience. *Journal of personality and social psychology*, 39(6): 1125.
9. Fukui, A.; Park, D. H.; Yang, D.; Rohrbach, A.; Darrell, T.; and Rohrbach, M. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.
10. Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 6–15.
11. Han, W.; Chen, H.; and Poria, S. 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*, 9180–9192.
12. Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, 2122. NIH Public Access.
13. Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. *MISA: Modality-Invariant and-Specific Representations for Multimodal Sentiment Analysis*, 1122–1131. Association for Computing Machinery.
14. He, X.; Du, X.; Wang, X.; Tian, F.; Tang, J.; and Chua, T.-S. 2018. Outer product-based neural collaborative filtering. *arXiv preprint arXiv:1808.03912*.
15. Hou, M.; Tang, J.; Zhang, J.; Kong, W.; and Zhao, Q. 2019. Deep multimodal multilinear fusion with high-order polynomial pooling. *Advances in Neural Information Processing Systems*, 32: 12136–12145.

 doi:10.20944/preprints202504.0397.v1

16. Hu, A.; and Flaxman, S. 2018. *Multimodal Sentiment Analysis To Explore the Structure of Emotions*, 350–358. New York, NY, USA: Association for Computing Machinery.
17. Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.
18. Liang, P. P.; Liu, Z.; Bagher Zadeh, A.; and Morency, L.-P. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *Conference on Empirical Methods in Natural Language Processing*, 150–161. Brussels, Belgium: Association for Computational Linguistics.
19. Lin, T.-Y.; RoyChowdhury, A.; and Maji, S. 2015. Bilinear cnn models for fine-grained visual recognition. In *Proceedings of the IEEE international conference on computer vision*, 1449–1457.
20. Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064*.
21. Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.
22. Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.
23. Mai, S.; Hu, H.; and Xing, S. 2020. Modality to Modality Translation: An Adversarial Representation Learning and Graph Fusion Network for Multimodal Fusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01): 164–172.
24. Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing)*, 1532–1543.
25. Pham, H.; Liang, P. P.; Manzini, T.; Morency, L.-P.; and Póczos, B. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6892–6899.
26. Pham, N.-Q.; Niehues, J.; Ha, T.-L.; and Waibel, A. 2019. Improving Zero-shot Translation with Language-Independent Constraints. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 13–23.
27. Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Annual Meeting of the Association for Computational Linguistics*, 873–883. Vancouver, Canada.
28. Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.
29. Rahman, W.; Hasan, M. K.; Lee, S.; Zadeh, A.; Mao, C.; Morency, L.-P.; and Hoque, E. 2020. Integrating multimodal information in large pretrained transformers. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2020, 2359. NIH Public Access.
30. Rajagopalan, S. S.; Morency, L.-P.; Baltrusaitis, T.; and Goecke, R. 2016. Extending long short-term memory for multi-view structured learning. In *European Conference on Computer Vision*, 338–353. Springer.
31. Sahu, G.; and Vechtomova, O. 2021. Adaptive Fusion Techniques for Multimodal Data. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3156–3166.
32. Siriwardhana, S.; Reis, A.; Weerasekera, R.; and Nanayakkara, S. 2020. Jointly Fine-Tuning "BERT-Like" Self Supervised Models to Improve Multimodal Speech Emotion Recognition. *Proc. Interspeech 2020*, 3755–3759.
33. Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. *arXiv e-prints*, 34: 8992–8999.
34. Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.
35. Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, 6558. NIH Public Access.
36. Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Annual Meeting of the Association for Computational Linguistics*, 6558–6569. Florence, Italy: Association for Computational Linguistics.
37. Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2019. Learning Factorized Multimodal Representations. arXiv:1806.06176.

 doi:10.20944/preprints202504.0397.v1

38. Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7216–7223.
39. Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10790–10797.
40. Yuan, J.; Liberman, M.; et al. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America*, 123(5): 3878.
41. Yuan, Z.; Sun, S.; Duan, L.; Li, C.; and Xu, C. 2020. Adversarial Multimodal Network for Movie Story Question Answering. *IEEE Transactions on Multimedia*, PP(99): 1–1.
42. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.
43. Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.
44. Zadeh, A.; Liang, P. P.; Mazumder, N.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Memory fusion network for multi-view sequential learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, volume 32.
45. Zadeh, A.; Liang, P. P.; Poria, S.; Vij, P.; Cambria, E.; and Morency, L.-P. 2018. Multi-attention recurrent network for human communication comprehension. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
46. Zadeh, A.; Zellers, R.; Pincus, E.; and Morency, L.-P. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88.
47. acob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186.
48. Endri Kacupaj, Kuldeep Singh, Maria Maleshkova, and Jens Lehmann. 2022. An Answer Verbalization Dataset for Conversational Question Answerings over Knowledge Graphs. *arXiv preprint arXiv:2208.06734* (2022).
49. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469
50. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
51. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.
52. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In P*roceedings of the 58th Annual Meeting of the Association for Computational Linguistic*s. 7871–7880.
53. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.
54. Pierre Marion, Paweł Krzysztof Nowak, and Francesco Piccinno. 2021. Structured Context and High-Coverage Grammar for Conversational Question Answering over Knowledge Graphs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2021).
55. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, April 2010. ISSN 0942-4962. doi:10.1007/s00530-010-0182-0.
56. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
57. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.

 doi:10.20944/preprints202504.0397.v1

58. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
59. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
60. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
61. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
62. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
63. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
64. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
65. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
66. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
67. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
68. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
69. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
70. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
71. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.
72. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
73. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
74. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
75. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
76. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
77. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
78. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.

 doi:10.20944/preprints202504.0397.v1

79. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
80. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
81. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
82. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
83. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
84. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
85. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
86. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
87. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
88. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
89. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
90. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
91. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
92. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
93. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
94. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
95. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
96. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
97. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
98. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
99. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.

 doi:10.20944/preprints202504.0397.v1

100. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
101. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
102. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
103. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
104. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
105. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.
106. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
107. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
108. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
109. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
110. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
111. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
112. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
113. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
114. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
115. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.