Integrating Commonsense Reasoning into Graph-based Multimodal Emotion Recognition

Xinyi Xu

Pembroke ID: Xu5OSRP25

Date of Submission: 26/07/2025

This dissertation is my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as specified in the text. It does not exceed the agreed word limit.

Integrating Commonsense Reasoning into Graph-based Multimodal Emotion Recognition

Xinyi Xu
University of Washington
Seattle, USA
xinyix26@uw.edu

***Abstract*—This paper presents CSGraphSmile, a multimodal emotion recognition model that integrates COMET-based commonsense reasoning adpoted from COSMIC into graph-based modality fusion adopted from GraphSmile. CSGraphSmile is trained and evaluated on the Multimodal EmotionLines Dataset, a benchmark for multimodal emotion recognition in conversation. Our approach combines role-sensitive speaker embeddings with a gated fusion mechanism to merge textual features from RoBERTa and commonsense features generated by feeding COMET knowledge into the CommonsenseRNN module published in COSMIC. The fused textual-commonsense representation is then processed by GraphSmile's heterogeneous graph fusion, which constructs cross-modality graphs to propagate emotion and sentiment information. Finally, it classifies emotion, sentiment, and sentiment shift for each utterance. CSGraphSmile achieves an accuracy of 67.2% and a weighted F1-score of 66.2%, comparable to leading methods such as M2FNet and GraphSmile and outperforming COSMIC and SACL-LSTM.**

***Index Terms*—Emotion recognition in conversation, Multimodal learning, Commonsense knowledge integration, Graph neural networks, Multimodal fusion**

I. Introduction

Emotion Recognition in Conversations (ERC) is a popular research problem in Artificial Intelligence (AI). ERC improves user interactions in many applications, such as mental health assistants and empathetic chat bots. Human conversations are inherently multimodal and complex. They consist not only of spoken words but also tone of voice and facial expressions, which all contribute to the speaker's emotional state [1]. Recognizing emotion in this context requires modeling temporal context, speaker dynamics, and relationship between context, tone, and facial expression. In recent years, the research community has focused on multimodal ERC, which leverages combinations of text, audio, and visual modalities to classify emotions [2].

Multimodal learning is the foundation of multimodal ERC. The challenges of multimodal learning have been categorized into representation, translation, alignment, fusion, and co-learning [1]. Some key open problems include modality heterogeneity and fusion [3]. These insights suggest how important modality fusion is to ERC.

In this study, we use the Multimodal EmotionLines Dataset (MELD) [4], which extends the EmotionLines dataset [5] by adding audio and visual modalities, making it a benchmark in multimodal ERC research. MELD consists of multiparty conversations extracted from the TV series *Friends*, including synchronized text, visual, and audio information with speaker, emotion, and sentimentt annotations. The dataset's complexity and multimodality make it ideal for evaluating models that understand emotional dynamics in conversations.

We will focus on two representative models: COmmonSense knowledge for eMotion Identification in Conversations (COSMIC) [6] and GraphSmile [7]. COSMIC uses commonsense knowledge to infer latent speaker states such as intent, reaction, and effect, enabling better modeling of emotional transitions. However, COSMIC only uses text and lacks multimodal fusion. In contrast, GraphSmile uses heterogeneous graph structures to capture sentiment dynamics and interactions across modalities.

We extend COSMIC and GraphSmile with a fusion mechanism that combines RoBERTa [8]textual features with COMET [9] commonsense featutrd derived through the CommonsenseRNN module of COSMIC [6]. These fused features are then incorporated into the heterogeneous graph fusion module of GraphSmile [7]. Our model architecture builds directly on COSMIC and GraphSmile and our primary contribution is integrating their strengths through gated fusion of textual and commonsense signals, resulting in stronger emotional reasoning in multimodal conversation.

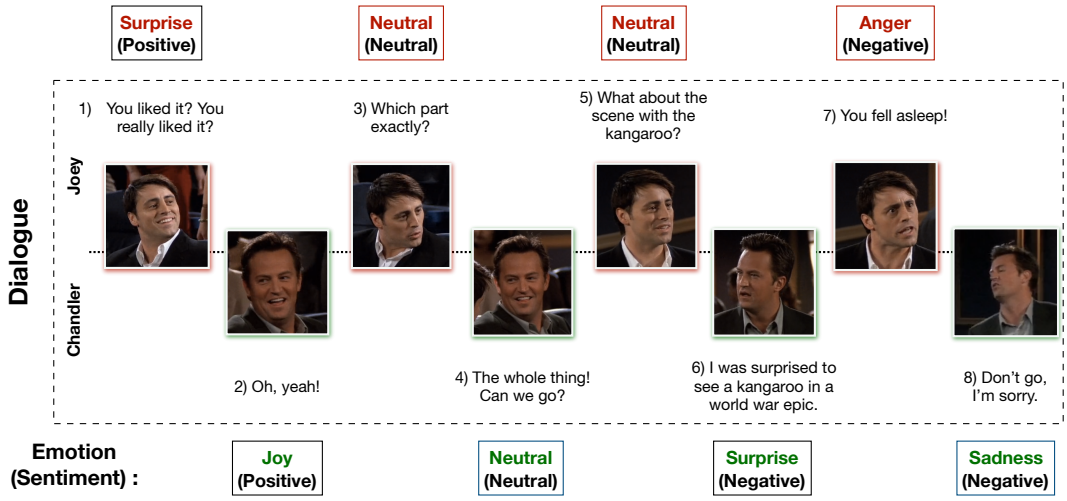


Fig. 1: Emotion shifts of speakers in a dialogue as compared to their previous emotions, copied from the MELD publication *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations* [4].

II. Literature Review

This section reviews the field of Emotion Recognition in Conversation (ERC), with a focus on multimodal approaches that uses text, audio, and video. We begin by introducing the significance and challenges of ERC, then a discussion on Multimodal Machine Learning that current state-of-the-art models are built on. We summarize several benchmark datasets in ERC, with a particular focus on Multimodal EmotionLines Dataset (MELD), the dataset used in our study. Next, we review leading models in the field, especially COSMIC and GraphSmile, which are the backbone of our proposed model.

A. Emotion Recognition in Conversation

Emotion Recognition in Conversation (ERC) is a specialized area in affective computing that aims to identify the emotional state of every utterance in a conversation, rather than treating utterances independently. ERC presents unique challenges such as modeling conversational context, speaker-specific dynamics, emotional shifts, and emotions such as sarcasm, which are absent in single-utterance emotion recognition tasks [10] [2]. For instance, an exclamation such as "Oh" alone may convey surprise, sarcasm, or disappointment depending on the preceding conversation. Therefore, ERC models must maintain awareness of conversation history and the role of the speaker and listener.

ERC has important real-world applications, such as empathetic conversational chat box, mental health assistant, educational tools [10]. For example, ERC can support health-care triage systems by flagging emotional distress. However, ERC also raises significant ethical concerns. Emotion recognition systems could put user autonomy and privacy in risk or exposed, especially when applied without informed consent [11]. There are also bias and fairness concerns, such that systems trained on limited or skewed datasets may misclassify emotions across demographic groups [12]. Additional concerns include emotional manipulation, surveillance, and the potential misuse of inferred emotional states in sensitive contexts such as mental health and education [13].

In summary, ERC advances human–machine interaction by enabling contextual, speaker-aware emotion detection, but there exists unique technical challenges and ethical risks that must be carefully managed.

B. Multimodality

In Multimodal Machine Learning, models are designed to interpret and integrate input from different modalities, such as text, audio, and video or image. Human perceptual experience is inherently multimodal, such that we see, hear, and feel our environment. Therefore, Artificial Intelligence (AI) should learn to jointly process these signals to achieve robust understanding mimicking human behaviors [1]. In the field of Multimodal Machine Learning, there are several core challenges, including learning representations that reconcile different data types, translating or aligning across modalities, fusing heterogeneous information, and enabling knowledge transfer when some modalities are scarce [1].

More recently, Multimodal Machine Learning is formalized into three key principles and six challenges. The principles are modality heterogeneity, similarity and shared concepts across modalities, and interactions between modalities. The challenges are representation,

TABLE I: Emotion and sentiment label distributions in the MELD dataset. Neutral utterances dominate the dataset (over 40%), while minority classes such as fear and disgust are less than 3% of the data.

Split	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Negative	Neutral	Positive	Total
Train	5180	1355	308	794	1906	293	1262	5180	2567	3351	11098
Test	1256	281	50	208	402	68	345	1256	521	833	2610

alignment, reasoning, generation, transference, and quantification [3].

Multimodal AI has applications in many domains. In healthcare, combined visual and textual data improve diagnostic accuracy. In robotics, integrating tactile, visual, and language feedback enhances autonomous decision-making. In embodied agents, coordinating video, language, and audio enables natural interaction with humans [3].

These taxonomies inspire our work in conversational emotion recognition (ERC). Emotions are conveyed through facial expressions, vocal tone, and textual context. When these modalities are fused, there is a richer understanding than any single input. Aligning these signals, modeling their interdependencies, and reasoning over their joint representation are well-known from multimodal theory [1] [3]. In this paper, we aim to apply these principles by integrating textual, acoustic, visual, and commonsense knowledge modalities to better interpret the emotional dynamics in conversations.

C. Emotion Recognition Datasets

Emotion Recognition in Conversation (ERC) relies on well-annotated datasets that capture dialogue structure, speaker roles, and often multiple modalities such as text, audio, and visual signals. This contrasts with single-utterance emotion tasks where there is contextual dependence throughout a conversation. We highlight three benchmark datasets below.

IEMOCAP (Interactive Emotional Dyadic Motion Capture database) [14] provides approximately 12 hours of conversations between professional actors, including text, audio, motion-capture video, facial expressions, and gestures. Utterances are annotated for six categorical emotions (happy, sad, neutral, angry, excited, frustrated) and continuous scores of valence, arousal, and dominance.

CMU-MOSEI (Multimodal Opinion Sentiment and Emotion Intensity) [15] provides over 65 hours of spoken monologue videos from 1,000+ speakers, with aligned text, audio, and visual modalities. Each segment includes annotations for six emotions (happy, sad, angry, fear, disgust, surprise), sentiment labels, and continuous emotion intensity scores.

MELD (Multimodal EmotionLines Dataset) [4] contains 13,708 utterances from 1,433 multi-party conversation from the *Friends* TV series, with aligned text, audio, and video. Each utterance is annotated with seven emotions (anger, disgust, fear, joy, neutral, sadness, surprise) and three sentiment classes (positive, neutral, negative).

We selected MELD for this study because of its realistic multi-party interactions and multimodal richness. The dataset's multi-speaker context challenges models to not only recognize static emotions but also track their evolution through both verbal and non-verbal interactions. By integrating commonsense reasoning, our approach aims to deal more effectively with minority emotion classes and leverage the complex interplay of modalities present in MELD.

D. Related Work

Emotion recognition models have evolved significantly over time. Early work focused on identifying emotions from single utterances. As conversational datasets are published, models began incorporating sequential context and speaker information in dialogues. More recently, multimodal ERC approaches have integrated audio and visual cues alongside text to capture more emotional signals.

Among contemporary ERC models, we summarize several of them with strong performance:

M3Net [16] builds graph neural networks over multimodal embeddings and introduces multi-frequency propagation to capture both local and global conversational context. It processes text with contextual Gated Recurrent Units (GRUs), and encodes audio and visual features via Multi-Layer Perceptrons (MLPs), aggregating them using a multi-frequency graph structure.

SACL-LSTM [17] uses sequence-based Long Short-Term Memory (LSTM) encoders augmented by supervised adversarial contrastive learning (SACL) and contextual adversarial training. This framework builds label-consistent representations and achieves top performance on several ERC datasets.

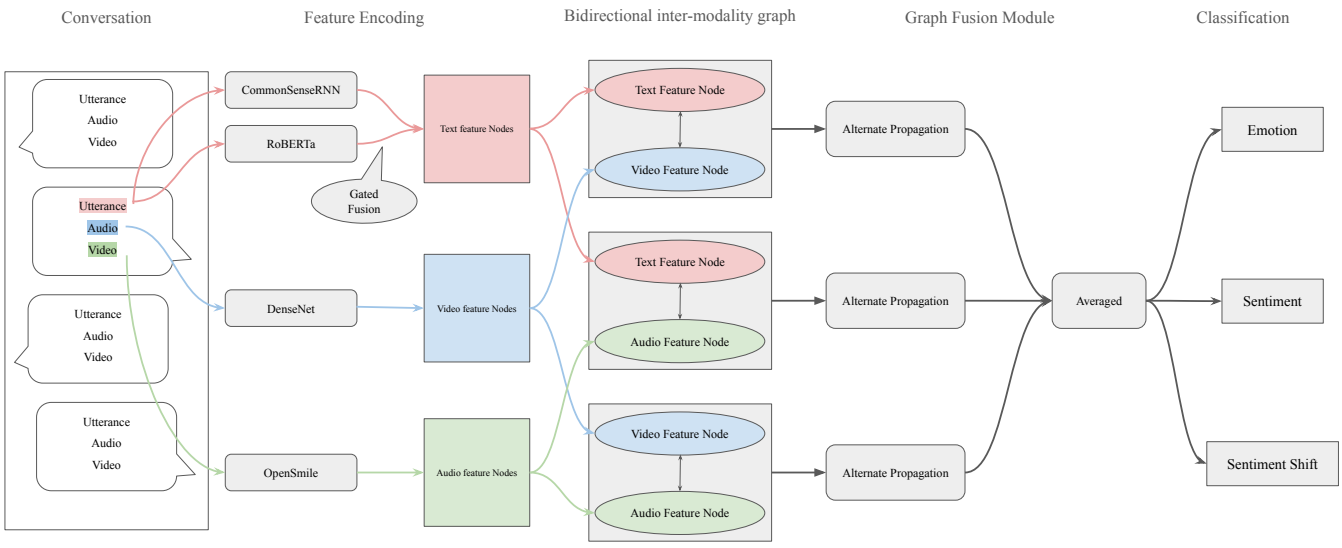


Fig. 2: Architecture of CSGraphSmile. The model fuses COMET-based commonsense features with RoBERTa-based textual features with gated fusion, and propagates the fused representations through heterogeneous multimodal graphs for emotion, sentiment, and sentiment shift prediction. The graph construction, graph fusion module, and output classification components are adapted from the model architecture figure from GraphSmile [7].

M2FNet [18] uses multi-head attention to fuse text, audio, and visual streams. It extracts each modality with tailored encoders and achieves top performance on IEMOCAP and MELD.

COSMIC (COmmonSense knowledge for eMotion Identification in Conversations) [6] greatly enhanced conversational emotion understanding by integrating commonsense knowledge into the model. It uses text embeddings from Commonsense Transformer (COMET) [9] for intents, reactions, and effects, feeding them into a text encoder and GRU modules. COSMIC excels at modeling latent speaker states and detecting emotion shifts. It explicitly separates and models internal speaker states, external listener states, and intent, making emotional dynamics more interpretable.

GraphSmile [7] captures modality relationships by constructing heterogeneous modality-pair graphs, and introduces the Graph Structure Fusion module to alternately propagate inter-modal and intra-modal cues through graph convolutions. It also performs a Sentiment Dynamics Perception task that explicitly identifies sentiment shifts between utterances. GraphSmile outperforms earlier models on benchmark datasets with its fusion between modalities.

COSMIC's commonsense guidance and GraphSmile's structured multi-modal fusion provides a robust foundations for our research. By combining COMET-informed commonsense embeddings with GraphSmile's graph fusion, we aim to capture both internal reasoning and inter-modal interactions in emotion shifts in conversations.

III. Methodology

This section describes the architecture and training procedure of our model, which integrates commonsense reasoning from COSMIC into multimodal graph fusion from GraphSmile. The model is trained and evaluated on the Multimodal EmotionLines Dataset (MELD) [4] to address three related subtasks: Emotion Recognition in Conversations (ERC), Sentiment Classification (SC), and Sentiment Shift Detection (SSD).

A. Dataset

MELD [4] extends EmotionLines [5] by including audio and visual channels. It consists of 1,433 multi-party dialogues with 13,708 utterances from the *Friends* TV Series. It is annotated for seven emotions (anger, disgust, fear, joy, neutral, sadness, surprise) and three sentiment labels. Specifically, given a dialogue as a sequence of utterances $u_i = [u_i^t, u_i^v, u_i^a]$ with speaker identities, where u_i^t, u_i^v, and u_i^a denote textual, visual, and acoustic modalities, the goal is to classify each utterance's emotion, sentiment, and determine whether a sentiment shift occurs relative to previous utterances. Figure 1 (copied from the MELD publication) illustrates how speaker emotions shift during a conversation, highlighting the dynamic interplay between text, facial expression, and speaker roles in emotional and sentiment classification.

Table I presents the class distributions across train and test splits. Validation split is randomly sampled from the train split (10%) in every training epoch. There exist significant imbalance such that over 40% of the train split is neutral utterances, while disgust and fear each represent only about 3%. Additionally, over half of the train split is negative in sentiment. Similar trends appear in the test split.

To mitigate class imbalance, we provide two types of balancing strategy, which are oversampling and subsampling. Oversampling replicates samples from minority classes until they reach a target size, which is set to be the smaller of twice the minimum class size or 30% of the total dataset size. This increases the number of samples from minority classes and avoids overfitting. Subsampling reduces the size of majority classes such that each emotion class is at most 1.5 times the minimum class size. For classes larger than this threshold, a random subset of samples is selected. This ensures that no class dominates the training data.

B. Feature Encoding

Due to time and equipment constraint, our model adopt precomputed features provided by prior work [7] [6] instead of performing feature extraction from raw modalities. Specifically, we use the released feature files from GraphSmile for textual, visual, and acoustic modalities, and the commonsense embeddings from COSMIC derived from COMET.

1) Textual Encoding: For text, we adopt the utterance-level RoBERTa embeddings released by GraphSmile. RoBERTa is a transformer-based language model optimized for robust pretraining on large corpora. These RoBERTa embeddings are 1024-dimensional vectors derived by averaging the final four hidden layers of RoBERTa's token representation [7] [8].

2) Commonsense Encoding: Commonsense features are adopted from COSMIC, which augments textual representations with inferred mental-state and causal knowledge. COSMIC leverages COMET [9], a transformer model trained on the ATOMIC commonsense knowledge graph, to predict if-then inferences about everyday events. ATOMIC organizes these inferences into nine relations including intent, needs, attributes, effects, and reactions of both the speaker and the listener [19].

Similar to COSMIC, we use five relation types that' are the most relevant to conversational emotion dynamics, which are intent of speaker, effect on speaker, reaction of speaker, effect on listener, and reaction of listener [6]. For each utterance, COMET generates embeddings corresponding to these relations by concatenating the utterance with each relation phrase and encoding them with transformer. The resulting five 768-dimensional vectors provide role-aware commonsense representations that capture inferred intentions, emotional reactions, and causal effects underlying the utterance.

3) Visual and Acoustic Encoding: For video and audio, we use GraphSmile's precomputed features. Visual features are extracted from video frames using DenseNet [20] pretrained on the FER+ dataset, which specializes in facial expression recognition. Then these video frames features are aggregated across each utterance to capture facial expressions that reflect emotions. Acoustic features are computed using the OpenSmile toolkit with the IS10 configuration [21], a widely adopted feature set in speech emotion recognition that encodes prosodic, spectral, and energy-related characteristics of speech [6]. These features summarize key signals such as pitch and intensity, which often correlated with emotions and sentiments.

C. Model Architecture

CSGraphSmile consists of two major components: a text-commonsense fusion module that combines RoBERTa-based textual features with COMET-based commonsense knowledge using a Gate Recurrent Unit (GRU) encoder and gating mechanism; and a heterogeneous graph structure fusion module that alternately propagates inter- and intra-modal information across three modality pairs using GraphSmile's graph convolution. Outputs from the graph modules are fused and passed to three classification heads, which are emotion recognition, sentiment classification, and sentiment shift detection.

1) RoBERTa-Commonsense Fusion: RoBERTa features and Commonsense Features are projected into a common hidden dimension using a linear layer with LeakyReLU activation. To incorporate speaker-specific information, we embed speaker roles and add them to the textual embeddings before fusion.

As shown in feature encoding in Figure 2, Commonsense Features are sent to the CommonsenseRNN module [6] to obtain Commonsense reasoning. This module models internal, external, and intent states of speakers and listeners using multiple GRU cells, guided by commonsense vectors from COMET representing five relation types: intent of speaker, effect on speaker, reaction of speaker, effect on listener, and reaction of listener. The GRU updates contextual hidden states sequentially across utterances to captures emotional dynamics of speaker and listener [6] [9].

TABLE II: Emotion classification on MELD test set. Neutral achieves the highest F1-score due to its dominance in the dataset, while fear and disgust perform worst, reflecting challenges with minority classes.

Label	Precision	Recall	F1-Score	Support
Neutral	0.**7573**	0.**8471**	0.**7997**	1256
Surprise	0.6255	0.5409	0.5802	281
Fear	0.2917	0.1400	0.1892	50
Sadness	0.5137	0.3606	0.4237	208
Joy	0.6930	0.5896	0.6371	402
Disgust	0.3958	0.2794	0.3276	68
Anger	0.4975	0.5797	0.5355	345
Weighted Avg	0.6611	0.6720	0.6621	2610
Overall Accuracy: 0.6720				

TABLE III: Sentiment classification on MELD test set. Negative sentiment achieves the highest F1-score and neutral sentiment is the worst, likely due to overlap with both positive and negative contexts.

Label	Precision	Recall	F1-Score	Support
Negative	0.**7693**	0.**8097**	0.**7890**	1256
Neutral	0.7575	0.5336	0.6261	521
Positive	0.6612	0.7311	0.6944	833
Weighted Avg	0.7324	0.7295	0.7263	2610
Overall Accuracy: 0.7295				

The output of CommonsenseRNN is then fused with RoBERTa features using a learnable gating mechanism, namely Gated Fusion in Figure 2. The gate g is parameterized as a three-layer multilayer perceptron (MLP) with nonlinear activations:

$$g = \sigma\big(W_3(\text{ReLU}(W_2(\text{ReLU}(W_1[\mathbf{x}_{\text{text}}\|\mathbf{x}_{\text{cs}}])))\big)$$

where $[\mathbf{x}_{\text{text}}\|\mathbf{x}_{\text{cs}}]$ denotes the concatenation of the RoBERTa features and output of CommonsenseRNN, W_1, W_2, W_3 are trainable weight matrices, ReLU introduces non-linearity, and σ is the sigmoid function. The resulting gating vector $g \in (0,1)^d$ adaptively controls the contribution of RoBERTa versus output of CommonsenseRNN in each dimension:

$$\mathbf{x}_{\text{fused}} = g \odot \mathbf{x}_{\text{text}} + (1-g) \odot \mathbf{x}_{\text{cs}}$$

This design allows our model to dynamically emphasize textual and commonsense signals over the other depending on the conversation.

2) Heterogeneous Graph Fusion: We build upon the Graph Structure Fusion (GSF) framework proposed in GraphSmile to integrate multimodal information across text, video, and audio. The architecture is illustrated in Bidirectional inter-modality graph and Graph Fusion Module in Figure 2, which is adpoted from the model architecture figure from GraphSmile [7]. The model constructs three heterogeneous graphs each representing a pair of modalities, which are text–visual, text–acoustic, and visual–acoustic. In these graphs, utterances are represented as nodes, and directed edges exist between nodes of the same utterance from different modalities under a temporal sliding window. Such model is able to capture both local and temporal interactions between modalities during a dialogue. The edge weights are trainable thus model learns the importance of information between modalities [7].

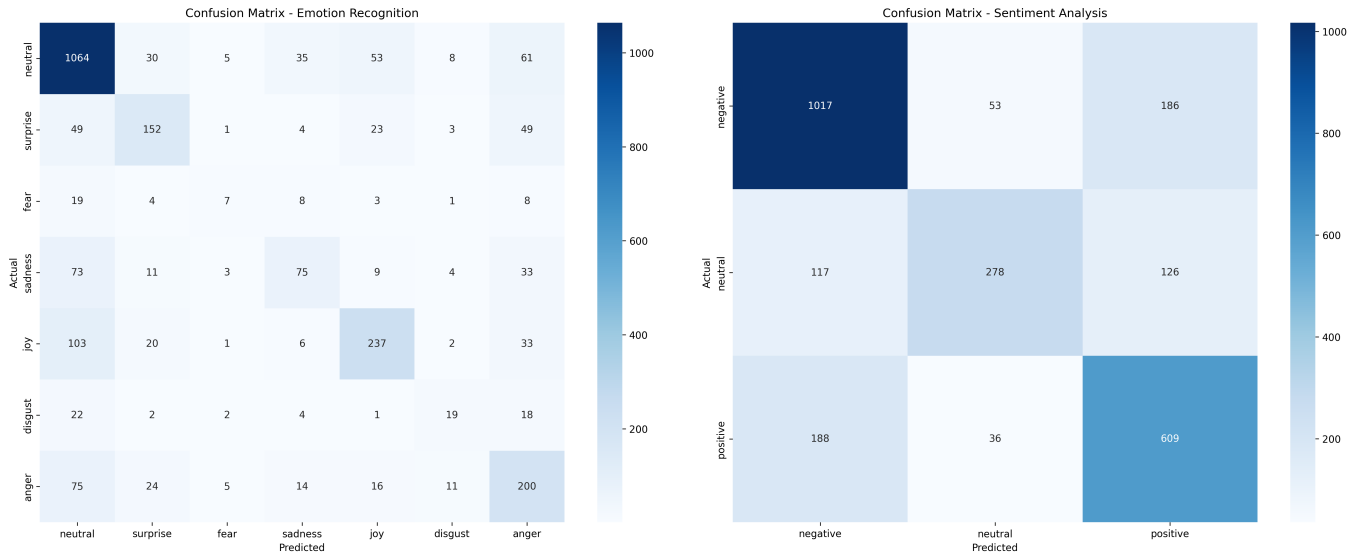
The GSF module uses alternate propagation, such that the first layer aggregates inter-modal cues, the second layer aggregates intra-modal cues, and so on. This prevents conflicts when combining multiple information sources simultaneously and ensures that both inter-modal dependencies and intra-modal contextual cues are captured effectively [7].

The outputs from the three heterogeneous graphs are then combined to form a fused multimodal representation, as shown in Figure 2. Each graph produces two directional outputs (for example, text influenced by visual features and visual influenced by text), resulting in six directional embeddings. These embeddings are projected into a common space and averaged such that all cross-modal interactions contributes equally [7]. The resulting fused multimodal representation integrates emotion cues from text, video, and audio, enhanced by the commonsense reasoning introduced earlier, and suppports later emotion, sentiment, and sentiment-shift predictions.

3) Output: The final fused multimodal representation is fed into three heads, each for a different task:

- Emotion Recognition: A linear layer with softmax to classify one of seven emotion labels for each utterance.
- Sentiment Classification: A parallel linear-softmax head to classify one of three sentiment labels.
- Sentiment Shift Detection: The SenShift_Feat module computes pairwise concatenations of utterance embeddings within a segment to predict sentiment shift labels (binary) between utterance pairs [7].

These three objectives are jointly optimized with a weighted sum of cross-entropy losses for emotion, sen-



(a) Emotion Classification Confusion Matrix

(b) Sentiment Classification Confusion Matrix

Fig. 3: Confusion matrices of classification results on the MELD test set. For emotion classification (a), neutral is predicted most accurately, while fear and disgust are often misclassified and confused with neutral or anger. For sentiment classification (b), negative sentiment dominates correct predictions, whereas neutral sentiment is frequently misclassified as either negative or positive.

timent, and sentiment shift, along with L2 regularization [7].

IV. Results

A. Training Setup

Our model is trained using the Adam optimizer, a learning rate of 7×10^{-5}, an L2 weight decay of 10^{-4}, and a batch size of 16 for 50 epochs on a single GPU. The Graph Structure Fusion module uses five heterogeneous graph convolution layers for each modality pair and a sliding window of size 3 for both past and future context. The sentiment shift detection window is also set to 3.

We adopt a multi-task objective that jointly optimizes emotion recognition, sentiment classification, and sentiment shift detection losses from GraphSmile [7]. The weights for the tasks (optimizing emotion recognition, sentiment classification, and sentiment shift detection losses) are set to 1.0, 0.5, and 0.2 based on validation performance.

B. Classification Results

In this subsection, we present the classification results of our final model on the MELD test set for both emotion and sentiment. Table II summarizes our final model's performance of emotion recognition and Table III summarizes the performance of sentiment classification. Precision measures how many utterances classified for a given class are actually correct. Recall measures how many true utterances of a class are successfully retrieved. F1-score is the harmonic mean of precision and recall. Support is the number of true instances of each class in the test set. Finally, we report weighted average of each metric across different labels and an overall accuracy of classification. Then we further analyze the confusion matrices to highlight common misclassifications patterns and discuss how class distribution impacts model performance, particularly for minority emotions.

1) Emotion Classification: From Table II, we see a weighted F1-score of 0.66 and an overall accuracy of 0.67. Among the seven emotion classes, neutral achieves the highest F1-score (0.80) with strong precision (0.76) and recall (0.85). This result is expected since neutral is the majority class in MELD (over 40% of the dataset) and provides abundant training samples leading to stable predictions. In contrast, fear shows the lowest F1-score (0.19), primarily due to its extremely low recall (0.14) and small support (only 50 samples). This reflects the challenge of learning rare classes in imbalanced datasets. Similarly, disgust and sadness also underperform due to limited number of samples in training data.

Joy and surprise achieve moderate F1-scores (0.64 and

TABLE IV: Ablation study with emotion classification F1-scores. The table demonstrates the performance of models from different development stages, starting from simple concatenation of features (Concat13) to the final CSGraphSmile configuration using selected RoBERTa layers {0,1}. Results show that gated fusion improves performance over dual encoders, while the final model achieves the highest weighted F1-score and overall accuracy.

Model Variants	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Accuracy	Weighted F1
Concat13	0.789	0.566	0.171	0.376	0.613	0.275	0.505	0.644	0.641
Dual Encoder	**0.799**	0.577	0.184	0.391	0.641	0.309	0.518	0.670	0.656
Dual Encoder Gated Fusion	0.797	0.585	**0.217**	0.399	0.636	**0.350**	0.534	0.670	0.660
CSGraphSmile	**0.799**	0.580	0.192	0.405	**0.644**	0.236	0.530	**0.672**	0.659
CSGraphSmile Subsampling	0.793	0.591	0.177	0.398	0.641	0.145	0.499	0.661	0.649
CSGraphSmile Oversampling	**0.799**	**0.585**	0.175	0.389	0.641	0.312	**0.539**	0.671	0.660
CSGraphSmile RoBERTa Layers 0,1	0.780	0.580	0.189	**0.424**	0.637	0.328	0.536	**0.672**	**0.662**

0.58 respectively). Although these emotions are better represented than fear or disgust, they are still misclassified as neutral and anger as shown in the confusion matrix in Figure 3(a). Overall, non-negative emotions such as neutral, joy, and surprise presents a higher F1-Score.

2) Sentiment Classification: From Table III, we see a weighted F1-score of 0.73 and accuracy of 0.73. Negative sentiment is most accurately classified with a F1-score 0.79 due to its large support (1256 samples). Positive sentiment has an F1-score of 0.69 and Neutral sentiment has an F1-score of 0.63. Similar trend appears in Figure 3(b), such that negative has a better performance than positive and positive has a better performance than neutral. Overall, sentiment classification has a better performance than emotion classification.

C. Ablation Study

To understand the contribution of each design choice in our model, we conduct ablation study on CSGraphSmile of different development stage, from simple feature concatenation to the fully fusion. Table IV presents per-class F1-scores, overall accuracy, and weighted F1-scores for each variant.

Concat13 is the baseline and simply concatenates RoBERTa features (from GraphSmile) and COMET commonsense features (from COSMIC). Without any fusion, this variant achieves a weighted F1-score of 0.641. Performance is reasonable for majority classes such as neutral but poor for minority classes like fear and disgust, which suggests model did not learn anything from commonsense knowledge and it might even confused the model.

Dual Encoder improves upon Concat13 by using separate encoders for RoBERTa and COMET features. This results in an improvement in weighted F1 to 0.656 and in accuracy to 0.670. Minority classes such as disgust and sadness benefited from it, which suggests separate encoding mitigates representational conflicts between RoBERTa and commonsense features.

Dual Encoder Gated Fusion uses a single-layer gating mechanism to weight textual versus commonsense signals for each utterance. This results in further gains in minority classes and improves weighted F1 to 0.660. The improvement suggests weighting helps capture when commonsense reasoning is more informative than RoBERTa features.

CSGraphSmile fully integrates CommonsenseRNN from COSMIC with GraphSmile's graph fusion. It uses multi-GRU reasoning over commonsense relations and fuses it with RoBERTa features using a deeper three-layer MLP gate. It achieves balanced performance across emotion classes, reaching 0.672 accuracy and 0.659 weighted F1, with improvements in joy and anger.

We also explored dataset balancing: **CSGraphSmile Subsampling** trims majority-class examples to reduce skew, while **CSGraphSmile Oversampling** augments minority classes. Oversampling yields better minority-class F1, but neither strategy significantly improves overall weighted F1.

Finally, **CSGraphSmile RoBERTa Layers 0,1** uses only the first two RoBERTa layers rather than all four, which results in the highest weighted F1 (0.662) and overall accuracy (0.672). We adopt this configuration as our final model because it balances performance across majority and minority classes while beating all other variants in weighted F1 and overall Accuracy.

D. Comparing with other Models

Table V compares our model, CSGraphSmile RoBERTa Layers {0,1}, with a few state-of-the-art models with performance reported on MELD emotion classification task. Overall, our model achieves an accuracy of 0.672 and a weighted F1-score of 0.662,

TABLE V: Comparison with representative models on MELD emotion classification. Our model achieves competitive performance, outperforming COSMIC and SACL-LSTM and closely matching GraphSmile and M2FNet. Results for representative models are reported from their original papers [16] [17] [18] [7] [6].

Models	Neutral	Surprise	Fear	Sadness	Joy	Disgust	Anger	Accuracy	Weighted F1
M3Net	0.791	0.595	0.133	0.429	0.651	0.217	0.535	0.666	0.658
SACL-LSTM	0.774	0.585	0.204	0.396	0.628	0.247	0.521	0.645	0.646
M2FNet	0.801	0.587	0.345	0.470	0.655	0.252	0.553	0.679	0.667
GraphSmile	0.804	0.591	0.182	0.425	0.650	0.324	0.537	0.677	0.667
COSMIC	-	-	-	-	-	-	-	-	0.652
CSGraphSmile RoBERTa Layers 0,1	0.780	0.580	0.189	0.424	0.637	0.328	0.536	0.672	0.662

which is competitive to M2FNet and GraphSmile. Our model beats SACL-LSTM and COSMIC. Notably, our model beats all others on the minority class disgust. The results shows integrating commonsense reasoning into graph-based multimodal fusion yields comparable performance to the best existing models, and we believe the performance could be further improved by extracting optimal features from raw MELD dataset instead of using existing ones published by GraphSmile and COSMIC.

Table V compares our model, CSGraphSmile RoBERTa Layers 0,1, against several representative models previously evaluated on the MELD dataset, including M3Net [16], SACL-LSTM [17], M2FNet [18], GraphSmile [7], and COSMIC [6]. Our model achieves an overall accuracy of 0.672 and a weighted F1-score of 0.662, which is competitive to M2FNet (0.667 Weighted F1) and GraphSmile (0.667 Weighted F1). This performance tells us that incorporating commonsense reasoning into multimodal graph fusion does not sacrifice predictive capability but complements it.

Table V also compares per-class F1 scores for those that are available. Our model performs particularly well on disgust (0.328), which is a minority emotion class in MELD such that it's only about 3% of all utterances. The integration of COMET-derived commonsense features likely contributes to this improvement, such that disgust requires reasoning about subtle speaker intentions and reactions. For neutral, the majority class, our model performs slightly lower than GraphSmile, which is potentially the trade-off between optimizing for majority and minority classes.

Compared to COSMIC, which only uses text inputs and reports reports a weighted F1 of 0.652, our model has a better performance across most metrics. Similarly, we beat SACL-LSTM (0.646 Weighted F1). These results highlight the contribution of integrating commonsense reasoning into graph-based fusion. We expect further improvements if future work replaces precomputed features with embeddings extracted directly from raw MELD data, ensuring better modality alignment and optimal performance.

V. Conclusion

This paper presents CSGraphSmile, a model that is built on two representative models in emotion recognition: commonsense reasoning from COSMIC [6] and heterogeneous graph multimodal fusion from GraphSmile [7]. Our contribution is bridging these two models using a gated fusion mechanism that integrates COMET-based [9] commonsense features with RoBERTa-based [8] textual embeddings before propagating them in multimodal graphs. We aim to capture mental-state cues from commonsense knowledge and cross-modal interactions between text, audio, and video.

Our experiments on the MELD dataset provides several findings. Incorporating commonsense features improves the classification of minority emotions such as disgust and anger. The gated fusion strategy provides a balance between model complexity and performance, resulting in the highest weighted F1-score and overall accuracy in all model variants. Ablation studies show that adding commonsense reasoning, gating, and RoBERTa layer selection each contribute improvements. Confusion matrix reminds us the challenge of imbalanced datasets, with neutral dominating predictions and minority emotions such as fear being misclassified. Although our model does not beat the strongest baselines like M2FNet and GraphSmile, it remains competitive and outperforms COSMIC and SACL-LSTM, especially in minority-class performance.

A major limitation is that we relied on precomputed features released by COSMIC and GraphSmile rather than extracting features from raw data. These features may not be optimal for our combined pipeline, resulting in non-optimal performance. In future work, we plan

to implement feature extraction to better align modality representations and commonsense reasoning. We will also address data imbalance using focal loss and data augmentation and extend our evaluation to other multimodal conversational datasets such as IEMOCAP and CMU-MOSEI for better generalization.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, 2019.

[2] P. Pereira *et al.*, "Deep emotion recognition in textual conversations: A survey," *arXiv preprint arXiv:2211.09172*, 2022.

[3] J. Li, Z. Gan, L. Wang, Z. Liu, and Z. S. Liu, "Foundations and trends in multimodal machine learning: Principles, challenges, and open questions," *Foundations and Trends in Machine Learning*, vol. 17, no. 1, pp. 1–203, 2024.

[4] S. Poria, D. Hazarika, N. Majumder, G. Naik, and E. Cambria, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[5] S.-Y. Chen, C.-C. Hsu, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "Emotionlines: An emotion corpus of multi-party conversations," in *Proc. 11th Int. Conf. Language Resources and Evaluation (LREC)*, Miyazaki, Japan, 2018.

[6] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Cosmic: Commonsense knowledge for emotion identification in conversations," in *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP)*, 2020, pp. 2470–2481.

[7] Y. Zhou, M. Wang, Y. Wang, and B. Wang, "Graphsmile: A lightweight and interpretable graph-based model for multimodal emotion recognition," *arXiv preprint arXiv:2407.21536*, 2024.

[8] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[9] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "Comet: Commonsense transformers for automatic knowledge graph construction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 4762–4779.

[10] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy, "Emotion recognition in conversation: Research challenges, datasets, and recent advances," *IEEE Access*, vol. 7, pp. 100 943–100 953, May 2019.

[11] D. Barker, M. K. R. Tippireddy, A. Farhan, and B. Ahmed, "Ethical considerations in emotion recognition research," *Psychology International*, vol. 7, no. 2, p. 43, 2025.

[12] A. Katirai, "Ethical considerations in emotion recognition technologies: A review of the literature," *AI and Ethics*, vol. 4, no. 4, pp. 927–948, 2023.

[13] G. Alhussein, I. Ziogas, S. Saleem, and L. J. Hadjileontiadis, "Speech emotion recognition in conversations using artificial intelligence: A systematic review and meta-analysis," *Artificial Intelligence Review*, vol. 58, no. 7, p. 198, Apr 2025.

[14] C. Busso *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, 2018, pp. 2236–2246.

[16] F. Chen, J. Shao, S. Zhu, and H. T. Shen, "Multivariate, multi-frequency and multimodal: Rethinking graph neural networks for emotion recognition in conversation (m3net)," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 761–10 770.

[17] D. Hu, Y. Bao, L. Wei, W. Zhou, and S. Hu, "Supervised adversarial contrastive learning for emotion recognition in conversations," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 10 835–10 852.

[18] V. Chudasama, P. Kar, A. Gudmalwar, N. Shah, P. Wasnik, and N. Onoe, "M2fnet: Multi-modal fusion network for emotion recognition in conversation," in *CVPR Workshop on Multimedia Understanding through Learning Associations (MULA)*, 2022.

[19] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 33, 2019, pp. 3027–3035.

[20] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[21] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062–1087, 2011.