Exploring Cross-Modal Interactions in Unimodal and Multimodal Emotion Recognition: An Empirical Study

Quanqi Du, Loic De Langhe, Els Lefever, Véronique Hoste
Language and Translation Technology Team (LT3), Ghent University
{quanqi.du, loic.delanghe, els.lefever, veronique.hoste}@ugent.be

Abstract

Understanding how cross-modal interactions influence unimodal and multimodal emotion recognition remains an open question in multimodal affective computing. This study presents a systematic empirical investigation of how multimodal inputs affect both unimodal and multimodal emotion recognition performance. Using the UniC dataset, which provides modality-specific and global multimodal annotations across text, audio, and visual modalities, we conduct experiments based on the Tensor Fusion Network (TFN) under unimodal, bi-modal, and tri-modal configurations. Results show that cross-modal interactions exert complex and asymmetric effects. While additional modalities can provide complementary emotional cues, they may also introduce interference when signals diverge. Models continue to struggle with less frequent or extreme emotions such as *disgust*. Notably, multimodal embeddings combined with unimodal annotations outperform fully multimodal supervision in the same setup, highlighting the role of annotation consistency and cue reliability. These findings provide a systematic empirical validation of the long-assumed notions, demonstrating that cross-modal effects are not simply additive and highlighting the need for more interpretable multimodal fusion strategies.

Keywords: multimodal emotion recognition, modality-specific annotation, cross-modal interaction

1. Introduction

Human emotion is inherently multimodal (Pan et al., 2023; Junchi et al., 2025). In daily communication, emotional meaning is conveyed not only through verbal expressions but also through other channels, such as tone and facial movements (Wallbott and Scherer, 1986). The interaction of these modalities enables humans to perceive and interpret emotions more accurately and richly than from any single channel alone. Inspired by this, multimodal emotion recognition (MER) has become a central task in affective computing (Picard, 1997) and human-computer interaction (Zhang et al., 2024), aiming to jointly analyze cues from multiple single modalities, for example, text, audio, and vision, to model emotion more comprehensively (Shou et al., 2025; Du et al., 2025a).

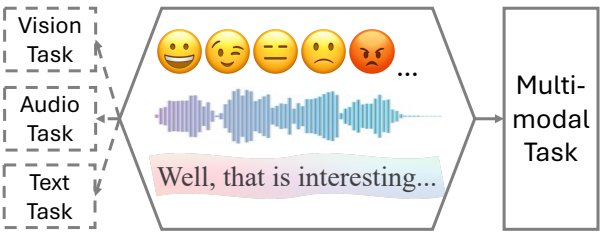


Figure 1: Interdependencies between modalities in emotion understanding

Over the past decades, extensive research has demonstrated that integrating multiple modalities improves emotion classification performance compared with unimodal baselines (Busso et al., 2004; Metallinou et al., 2010; Ezzameli and Mahersia, 2023). The performance gain is generally attributed to cross-modal interactions, where information from one modality helps disambiguate or reinforce cues from another (Yu et al., 2020). For instance, as shown in Figure 1, the same sentence "Well, this is interesting..." can express a whole range of emotions depending on tone of voice or facial expression. Modeling such interdependencies is thus crucial for robust emotion understanding.

However, while cross-modal interaction is known to benefit multimodal emotion classification (Kumar and Vepa, 2020), a key theoretical question remains underexplored: Can cross-modal interactions also influence unimodal tasks, i.e. settings in which a single modality's label is taken as the gold label? And how does this influence relate to the task performance of multimodal models and the way multimodal models represent and integrate information across modalities?

In other words, beyond improving fused outputs, does the integration of multiple modalities shape how individual modalities encode emotion? Addressing this question is crucial for understanding the representational dynamics between unimodal and multimodal processing. Human perception provides suggestive evidence that such influence exists – people's interpretation of speech or facial emotion is often modulated by contextual knowledge derived from other modalities (Barrett et al., 2011; Wieser and Brosch, 2012). Investigating whether this phenomenon also emerges in com-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 105–115
16 May 2026.

putational models can deepen our understanding of multimodal representation learning and affective reasoning.

To explore this issue, we propose a two-stage experimental framework that systematically investigates the interplay between unimodal and multimodal learning. In the first stage, we train a trimodal (Text-Audio-Video, TAV) model using multimodal annotations to capture integrated emotional representations. In the second stage, we conduct a series of controlled experiments with unimodal and bimodal configurations (T, A, V, TA, TV, AV), where each setting is trained on its corresponding unimodal annotations. For bimodal configurations such as TA, this means leveraging the annotations from both constituent unimodal modalities (i.e., T and A), respectively. This structure enables both vertical comparison between models trained and evaluated with unimodal and multimodal annotations, and horizontal comparison among unimodal, bimodal, and multimodal inputs.

Through this design, we aim to explore two fundamental questions:

- Do cross-modal interactions, introduced through the addition of other modalities, also influence unimodal tasks? This possibility is hypothesized, but experimental evidence to substantiate it is currently lacking.

- How are such influences related to the performance and interpretability of multimodal classification? This relationship remains insufficiently explored, and systematic empirical investigation is still lacking.

Our results reveal that cross-modal interactions indeed influence unimodal emotion prediction. In particular, we find that models trained on multimodal annotations learn richer, more balanced affective representations, shaping the final emotion representation, moving it beyond a simple addition of modalities. Moreover, unimodal tasks display distinct prediction tendencies that can be traced to patterns of cross-modal correlation observed during multimodal training. These findings suggest that the model's emotion recognition process operates along a continuum, where unimodal and multimodal processing are interdependent rather than isolated.

By analyzing these phenomena empirically, this work presents a dedicated investigation into how cross-modal interactions exert positive or negative influences on unimodal and multimodal emotion tasks. It thereby advances a deeper understanding of cross-modal dynamics and provides a novel interpretative perspective on the multimodal emotion research.

2. Related Work

2.1. Unimodal and Multimodal Emotion Recognition

Emotion recognition has long been a central topic in affective computing (Picard, 1997). Early approaches primarily relied on unimodal cues, such as text (Stajner and Klinger, 2023), speech (Slaughter et al., 2023), facial expressions (Leong et al., 2023), to infer affective states from a single information channel. However, human emotion expression is inherently multimodal (Pan et al., 2023; Junchi et al., 2025), where the same verbal content may convey drastically different emotions depending on tone or facial gestures. This realization has led to a paradigm shift toward MER, which seeks to integrate different signals from multiple modalities for more robust affective inference (Zhang et al., 2024; Hazmoune and Bougamouza, 2024).

Recent advances in deep learning have greatly enhanced MER performance by enabling the joint learning of hierarchical and correlated representations. Representative models include graph-based fusion networks (Li et al., 2023), transformer-based learning frameworks (Khan et al., 2025), and contrastive learning strategies (Xie et al., 2025). By employing graph-based relational modeling, self-attention mechanisms, and contrastive representation learning, these architectures effectively capture inter- and intra-modal dependencies, achieving notable gains over traditional baselines.

Despite this progress, most multimodal systems are designed primarily to maximize the overall predictive performance, rather than to investigate how modalities interact internally during learning.

2.2. Modeling Cross-Modal Interactions

Understanding cross-modal interaction mechanisms is crucial for interpreting multimodal learning (Fu et al., 2024). Researchers have explored various methods to model cross-modal interactions, either implicitly or explicitly. For instance, MulT (Tsai et al., 2019) employs directional pairwise cross-modal attention to capture interactions between multimodal sequences, while CENet (Wang et al., 2023) enhances text representations by integrating visual and acoustic information into a language model. Additionally, LDW-MTFN (Du et al., 2025a) introduces a label-based weighting scheme to explicitly balance the contributions of different modalities.

However, existing mechanisms are primarily designed to improve overall multimodal performance and are less informative for understanding naive cross-modal interactions. They tend to select only the most useful information from each modality by using mechanisms such as cross-modal attention

(Tsai et al., 2019) and gating (Sun et al., 2024), rather than considering all available information. Moreover, previous studies have examined how cross-modal interactions affect multimodal task outcomes, but not their impact on unimodal tasks in the field of emotion recognition, as illustrated in Figure 1. Therefore, to investigate authentic cross-modal interactions, it is necessary to fully integrate information from all modalities without any selective filtering, while also paying attention to how these interactions influence unimodal tasks.

2.3. Dataset-Level Advances and Remaining Gaps

One major bottleneck in examining cross-modal effects lies in the lack of datasets with modality-specific annotations. Most existing MER datasets, such as MSP-IMPROV (Busso et al., 2016) or CMU-MOSEI (Zadeh et al., 2018), only provide unified emotion labels derived from the overall multimodal impression. This design hinders a fine-grained comparison between unimodal and multimodal emotion detection performances, as each modality's individual contribution cannot be independently assessed.

To address this limitation, Yu et al. (2020) introduced the Chinese dataset CH-SIMS, annotated at both unimodal and multimodal levels. Subsequently, Liu et al. (2022) expanded this resource by releasing CH-SIMS V2, adding more instances to the original dataset. Both versions of the dataset are annotated with sentiment labels. More recently, Du et al. (2025b) proposed another dataset, UniC, with both sentiment and categorical emotion labels in English. An notable difference between these two datasets is that CH-SIMS contains acted emotional expressions while the UniC dataset comprises authentic emotional expressions. Both datasets provide annotations for text, audio, and visual modalities.

Our study focuses exclusively on the UniC dataset, as it captures spontaneous and authentic emotional expressions, making it more suitable for robust emotion analysis. Leveraging its categorical emotion labels on both unimodal and multimodal level, we train and evaluate models under unimodal, bimodal, and trimodal configurations. This setup provides a rare opportunity to investigate how cross-modal interactions influence unimodal and multimodal emotion recognition, addressing an important gap in existing emotion research.

3. Method

3.1. Experimental Framework

To address the research question of whether cross-modal interactions can influence unimodal tasks and how this relates to multimodal classification, we design a two-stage experimental framework. In the first stage, a trimodal configuration of text, audio and video (TAV) is trained and evaluated using multimodal annotations to capture integrated emotional representations. In the second stage, experiments are conducted with unimodal or bimodal inputs (T, A, V, TA, TV, AV), where each setting is trained with the corresponding unimodal annotations. This design enables a systematic comparison between multimodal and unimodal learning, providing insights into how cross-modal information contributes to unimodal emotion prediction and overall multimodal understanding.

3.2. Dataset

We conduct our experiments on the UniC dataset (Du et al., 2025b), which contains 965 video clips featuring genuine emotional expressions. UniC was annotated by three trained college students from Ghent University, who served as experts, with inter-annotator agreement used to ensure annotation quality. Unlike acted emotion datasets such as CH-SIMS (Yu et al., 2020), UniC captures authentic emotions, making recognition both more challenging and more representative of real-world scenarios. Importantly, besides an overall emotion annotation, UniC also provides independent categorical emotion labels for three single modalities: text, audio, and (silent) video. To the best of our knowledge, it is the only available dataset that offers modality-specific categorical annotations for emotions, thus enabling a systematic study of unimodal baselines, cross-modal interactions, and label-specific emotion classification tasks.

UniC contains both dimensional and categorical emotion labels. For the former, valence and arousal are evaluated as an integer from 1 to 5, while for the latter, a set of seven categorical emotion labels are used, including *disgust*, *disappointment*, *confusion*, *neutral*, *surprise*, *contentment* and *joy*.

The UniC dataset releases only pre-extracted feature packages rather than the raw video data, primarily due to privacy and data protection considerations. Specifically, the released features are extracted using multilingual BERT-base model (Devlin et al., 2019) for text, openSMILE (Eyben et al., 2010) for audio, and MediaPipe (Lugaresi et al., 2019) (25 FPS) for the visual modality.

3.3. Model

To investigate cross-modal interactions, it is essential to select a model that can faithfully capture the contributions of each modality. Many existing multimodal models, such as MulT (Tsai et al., 2019), CENet (Wang et al., 2023), and LDW-MTFN (Du

et al., 2025a), incorporate mechanisms like gating, dynamic weighting, or cross-modal attention. While these mechanisms often lead to performance improvements, they complicate the understanding. The observed improvements may primarily reflect architectural biases rather than the genuine effect of modality interactions, as such mechanisms selectively emphasize certain inputs and suppress others, potentially masking the natural influence of unimodal signals.

In contrast, the Tensor Fusion Network (TFN) (Zadeh et al., 2017) provides a static and relatively exhaustive representation of cross-modal interactions, without relying on any of the aforementioned architectural changes. Formally, let x_T, x_A, and x_V denote the embeddings of text, audio, and video modalities, respectively. The tensor fusion representation is computed as

$$\mathcal{T} = \begin{bmatrix} x_{\mathcal{T}} \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_A \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_V \\ 1 \end{bmatrix} \quad (1)$$

where $\otimes$ denotes the outer product. The resulting tensor $\mathcal{T}$ is then flattened and passed through fully connected layers for classification. This approach preserves all information from individual modalities while capturing deep pairwise and higher-order interactions, unlike simple concatenation, which does not fully exploit the information content. Therefore, TFN is particularly suitable for studying the genuine contributions of added modalities, as any observed performance changes can be more directly attributed to the modalities themselves rather than to architectural artifacts.

Although multi-task models such as MTFN have shown stronger performance in prior studies, we do not adopt a multi-task approach in this work, as it could introduce confounding factors that obscure the direct effects of each modality.

Overall, TFN provides a clear and controlled framework for studying intrinsic cross-modal interactions, making it a natural choice for our analysis.

3.4. Implementation Details

We define three input modalities: text (T), represented by textual transcript embeddings; audio (A), extracted from speech acoustic features excluding transcripts; and video (V), extracted from silent video frames. To investigate cross-modality interaction effects, we consider all possible modality combinations: unimodal (T, A, V), bimodal (TA, TV, AV), and trimodal (TAV). Each configuration serves as input to the model. Independent unimodal and multimodal annotations in UniC enable the model to be trained and evaluated for different tasks, facilitating a systematic analysis of cross-modal influences. The UniC dataset is partitioned into training, validation, and test subsets following a 6:2:2 ratio.

All models are trained using cross-entropy loss for seven categorical emotion labels, with a batch size of 32 and a learning rate of 0.0005. The experiments are carried out on NVIDIA Tesla V100-SXM2-16GB GPUs, and repeated with three random seeds to ensure robustness.

Performance is evaluated using accuracy and weighted F1-score to account for class imbalance, allowing us to establish robust unimodal baselines and quantify both positive and negative effects of cross-modal interactions.

4. General Results

4.1. Unimodal-label Settings

Table 1 summarizes the performance of TFN under different input modalities, evaluated against four distinct annotation sources (text-, audio-, and vision-based and multimodal labels).

First of all, it is clear that for unimodal emotion recognition tasks – that is, when using unimodal labels (text, audio and vision labels, respectively) – the model achieves the best performance when the three modalities are combined. This suggests that multimodal information fusion can provide complementary cues that enhances emotion recognition in single-modality tasks. Among the three unimodal-label settings, the model performs best with text labels, reaching an accuracy of 34.90% and an F1-score of 28.09%, followed by the vision setup (Acc=31.24%, F1 = 25.98%) and the audio setup (Acc=27.23%, F1=24.00%). These results indicate that when the three modalities are combined, the model finds it easier to learn text-related emotional patterns than those in audio or visual modalities.

When using text annotations, the unimodal text baseline (T) achieves a higher accuracy (33.86%) than both the text-audio (TA, acc=29.67%) and text-vision (TV, acc=32.46%) settings. However, it shows the lowest F1-score (19.13%), suggesting that while the text-only model can capture the majority class more frequently, it struggles to identify minority emotional categories. This suggests that the inclusion of additional modalities introduces beneficial complementary information that improves classification balance and robustness, even if overall accuracy does not always increase. A further investigation reveals that the unimodal text baseline predicted all instances as *neutral*, as shown in Figure 2 in Section 5, highlighting the limitation of the text-only setup. An inspection of the textual annotations in UniC shows that *neutral* is indeed the most frequent label, accounting for approximately one third of all instances, which may have reinforced this prediction bias in the text-only setup.

In contrast, the pattern is reversed when using audio annotations. The unimodal audio

Input	Text Label		Audio Label		Video Label		Multimodal	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
TAV	**34.90**	**28.09**	**27.23**	**24.00**	**31.24**	**25.98**	28.45	21.17
TA	29.67	20.30	23.91	18.88	–	–	27.22	19.57
TV	32.46	25.74	–	–	27.92	21.35	28.27	21.84
AV	–	–	23.21	20.78	31.24	25.36	27.57	18.12
T	33.86	19.13	–	–	–	–	28.10	16.99
A	–	–	25.83	22.47	–	–	**30.72**	**24.65**
V	–	–	–	–	28.80	22.70	28.80	18.77

Table 1: Performance of TFN when using text, audio, and video labels, respectively. *Input* refers to the combination of modality embeddings. Results are averaged from three experiments with different random seeds.

baseline (A) outperforms the multimodal settings, achieving both higher accuracy (25.83%) and F1-score (22.47%) than text-audio (TA, acc=23.91%, F1=18.88%) and audio-vision (AV, acc=23.21%, F1=20.78%). This indicates that the addition of text or visual features introduces interference or conflicting information, reducing the model's ability to align with the emotion expressed in the audio modality. On the other hand, we also found that during training, the inclusion of additional modalities also leads to substantially earlier convergence, indicating that they provide auxiliary information that helps the model learn more efficiently, even if the final performance does not always improve.

The situation becomes more complex when the model uses vision annotations. The unimodal vision baseline (V, Acc = 28.80%, F1 = 22.70%) performs better than the text-vision model (TV, Acc = 27.92%, F1 = 21.35%), but worse than the audio-vision model (AV, Acc = 31.24%, F1 = 25.36%). This pattern suggests that adding the audio modality provides positive complementary information, while adding text may have a negative or distracting influence on visual emotion recognition.

Taken together, across unimodal, bimodal, and trimodal setups for unimodal-label tasks, the results indicate that adding one additional modality can have either positive or negative effects, depending on the degree of alignment between modalities and label sources. However, when two additional modalities are combined, the overall effect tends to be positive, suggesting that the benefits of multimodal complementarity outweigh the potential interference among modalities.

4.2. Multimodal-label Settings

When the model is trained and evaluated with multimodal labels, however, the results exhibit a distinct pattern compared with the unimodal-label settings. As presented in Table 1, the model attains the highest accuracy (30.72%) and F1-score (24.65%) when using the audio modality alone, outperforming other unimodal settings and all multimodal combinations. A possible explanation for this result could be that the audio modality carries strong emotional cues that are most consistent with the integrated multimodal annotations, reflecting its critical role in conveying affective information such as tone, rhythm, and prosody.

In contrast, the performance of the full three-modality fusion (TAV) does not exceed that of the audio-only settings, although it still surpasses the video-only and text-only configurations in terms of F1-score (21.17% > 18.77% > 16.99%). This indicates that combining the three modalities provides certain complementary benefits, but may also introduce redundant or conflicting information.

4.3. Cross-setting Interpretation

When examining the first line of Table 1, it is evident that when text, audio, and vision are used together as inputs, the model performs better when trained and evaluated with unimodal annotations than with multimodal annotations. This observation suggests two possible explanations. First, some emotional cues emphasized in each unimodal annotation may not be fully consistent with the holistic perception reflected in the multimodal labels. Second, multimodal annotations, while more comprehensive, may introduce additional uncertainty or ambiguity, making it harder for the model to align integrated features with a unified emotional target.

It is also observed that the model achieves its best performance when using the trimodal input but being trained and evaluated with textual annotations, suggesting a clear textual predominance. This finding is consistent with previous studies, which have similarly reported that textual features tend to dominate in multimodal emotion recognition tasks (Liu et al., 2022).

Comparing the two experiment schemes, it becomes evident that the relationship between modality and annotation source critically determines model performance. When using unimodal annotations, multimodal fusion is beneficial because each label corresponds to its own modality, and

the integrated features complement one another. When using multimodal annotations, however, the target label represents an overall emotional judgment, which may not align equally across modalities. Consequently, modalities like audio, which might naturally reflect more global affective impressions, could obtain better performance than the other two single modalities. However, examining the bimodal setups reveals that combinations including text, specifically TA (text + audio) and TV (text + vision), still outperform AV (audio + vision). This suggests that textual predominance remains.

5. Qualitative Analysis

Figure 2 illustrates the confusion matrices across all modality configurations. The first three rows correspond to the different unimodal annotations used, namely, text-, audio-, and vision-based annotations, and the columns represent the gradual inclusion of modalities (single → bimodal → trimodal). The fourth row represents the single modal inputs, while the fifth refers to the bimodal and trimodal inputs. The last two rows both refer to the results of models using multimodal annotations. The seven emotion categories (disgust, disappointment, confusion, neutral, surprise, contentment and joy) are indexed as 0-6.

The first row in Figure 2 presents the confusion matrices for the text-based classification task under different modality settings (T, TA, TV, and TAV). In the purely textual condition (T-T), the model exclusively predicts *neural* (class 3) regardless of the ground-truth label, indicating a strong bias toward the dominant class and a lack of discriminative capacity when relying solely on textual cues. When additional modalities are introduced, this bias is gradually mitigated. In the T-TA and T-TV settings, although *neutral* remains the most frequent prediction, the model begins to recognize a small number of other categories (e.g., *disgust*, *disappointment*, and *contentment*, or classes 0, 1, 5), suggesting that audio and visual information help the model capture subtle emotional variations that are not easily inferred from text alone. The T-TAV configuration further enhances this diversification: predictions are more evenly distributed across classes, and off-diagonal entries (especially toward *neutral*, *contentment* and *joy*, or classes 3, 5 and 6) increase, reflecting improved sensitivity to inter-class distinctions. Nevertheless, confusion among neighboring emotional categories persists, for example, *contentment* and *joy* (class 5 and 6), implying that while multimodal cues alleviate the single-modality bias, cross-modal interactions remain imperfect. Notably, the emotion *joy* is predicted only when multimodal information is available.

Compared with the text-based experiments, the audio-based experiments show predictions that are more evenly distributed across all emotion categories except *confusion* and *surprise*, and remain more stable across different modality settings. The range and number of correctly predicted emotion categories do not fluctuate as drastically as in the text-based results, suggesting that audio-based emotion recognition is relatively more stable and less sensitive to cross-modal variations. However, it is also noticed that there is no *disgust* emotion prediction in the TA setup, suggesting that the addition of textual information may have biased the model toward more neutral or less extreme affective interpretations, thereby suppressing the recognition of *disgust* with distinctive acoustic patterns.

As for the vision-based experiments, the results show different patterns from the text- and audio-based setups. The unimodal setting (V-V) shows a roughly balanced number of correct predictions between *disappointment* (22) and *contentment* (21). When textual information is added (V-TV), the model becomes more sensitive to negative emotions, as *disappointment* increases to 29 while *contentment* drops to 14. In contrast, adding audio (V-AV) strengthens positive recognition: *contentment* rises to 26 and *joy* also improves notably from 7 to 16, whereas *disappointment* decreases to 19. In the full multimodal condition (V-TAV), the results appear to balance these trends, with *disappointment* and *contentment* reaching 25 and 19 respectively, suggesting that the joint contribution of text and audio helps the model achieve a more balanced perception between negative and positive emotions.

In the multimodal label experiments, the three unimodal settings (M-T, M-A, and M-V) show a clear tendency toward predicting more positive emotions overall. Among them, *contentment* is consistently the most correctly recognized category, exceeding *disappointment* in all cases. This suggests that when trained with multimodal annotations, the unimodal models tend to align more with the positive affective patterns emphasized in the multimodal ground truth. Compared to text and vision, the audio-based model exhibits a more balanced distribution across positive and negative categories and is also the only one that successfully predicts *joy*, highlighting the unique contribution of acoustic cues to the recognition of high-arousal positive emotions.

For the multimodal-label experiments involving two- and three-modality combinations, several patterns emerge. Bi-modal setups that include audio (M-TA, M-AV) inherit the audio modality's ability to predict *joy*, whereas the text+vision combination (M-TV) predicts two *joy* instances, but both are incorrect, indicating that audio is critical for high-arousal positive emotion recognition. The tri-modal

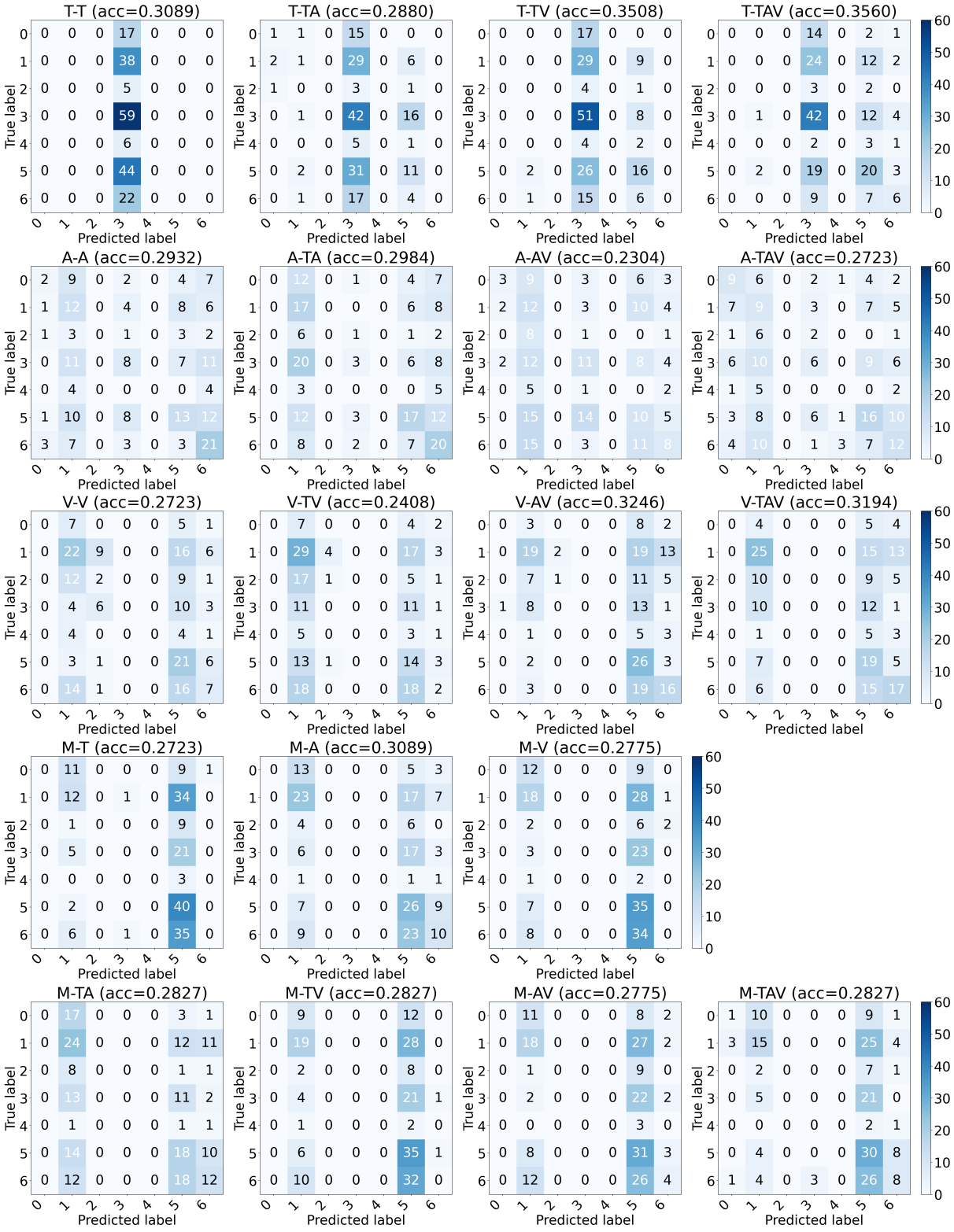


Figure 2: Confusion matrices of true and predicted labels, when the model is trained and evaluated using different annotation types and various modality combinations. T, A, V, and M refer to text, audio, vision, and multimodal setups, respectively. Labels 0-6 correspond to seven emotions: *disgust*, *disappointment*, *confusion*, *neutral*, *surprise*, *contentment*, and *joy*.

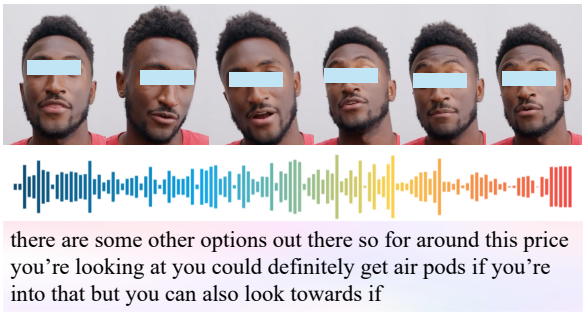
configuration (M-TAV) shows similar capabilities for *joy* as the bi-modal audio setups and is also the only setting among all multimodal-label experiments to produce predictions for *disgust*, with one correct out of five predicted instances. Additionally, the M-TAV results exhibit more frequent confusion between emotions of the same overall sentiment, particularly between *contentment* and *joy*, suggesting that while integrating three modalities balances the strengths of individual channels, it also introduces subtle intra-sentiment ambiguity in the model's predictions.

6. Discussion and Conclusion

This study, to the best of our knowledge, provides a systematic empirical analysis that examines how the presence of the different modalities influences emotion recognition when trained and evaluated under both unimodal and multimodal supervision.

6.1. Cross-Modal Complementarity and Interference

Our results reveal that cross-modal interactions exert a complex impact on both unimodal and multimodal emotion recognition tasks. Rather than simply enhancing recogniton accuracy through information aggregation, these interactions entail a trade-off between complementary and conflicting emotional cues. When using unimodal annotations, tri-modal configurations generally outperform unimodal and bi-modal ones, suggesting that richer multimodal representations can provide complementary evidence for emotion inference. However, the addition of a single auxiliary modality does not consistently yield improvements. This indicates that adding another modality can sometimes provide complementary information, but may also introduce interference or misalignment between modalities when their emotional signals diverge.



there are some other options out there so for around this price you're looking at you could definitely get air pods if you're into that but you can also look towards if

Figure 3: An instance from the UniC test dataset, labelled as *neutral*, *contentment*, *disgust*, and *neutral* in the text, audio, silent video, and multimodal setups.

In practice, multimodal data often contain heterogeneous and sometimes conflicting emotional expressions, such as discrepancies between facial expressions, vocal tone, and textual content, as shown in Figure 3. This inherent variability means that cross-modal interactions can amplify or diminish the contribution of each modality depending on the context, making the impact of adding modalities highly situation-dependent. Understanding and modeling this nuanced balance is therefore crucial for robust multimodal emotion recognition.

6.2. Supervision Type and Annotation Reliability

Interestingly, models trained and evaluated with unimodal annotations tend to outperform those relying on multimodal annotations, revealing a counterintuitive phenomenon in multimodal emotion recognition. A likely explanation lies in the annotation reliability: unimodal annotations, which focus on a single information channel, tend to be more internally consistent, whereas multimodal annotations often suffer from lower inter-annotator agreement (Du et al., 2023). This may occur because multimodal judgements may be influenced by differing weights assigned to the visual, acoustic, or linguistic cues. As a result, multimodal labels often introduce greater uncertainty. These observations suggest that combining multimodal input embeddings with unimodal annotations can sometimes yield more stable and effective results than relying solely on fully multimodal annotations.

6.3. Methodological Implications for Fusion Strategies

The present study also highlights the methodological implications of using Tensor Fusion Networks, or TFNs, which treat all modalities equally. While this approach can capture complementary information, it is suboptimal when large discrepancies exist across modality-specific annotations. Prior research has shown that weighing modality embeddings appropriately can improve performance with multimodal annotations (Du et al., 2025a). It remains, however, an open question whether a similar strategy applied to weighted embeddings with unimodal annotations would further enhance performance, and how it would compare to existing configurations. Exploring this possibility could provide valuable insights into optimizing multimodal emotion recognition under varying levels of annotation reliability.

6.4. Main Contributions

This paper makes three main contributions. First, it establishes a unified experimental framework for

directly comparing unimodal and multimodal supervision under identical modality configurations. Second, it clarifies that cross-modal interactions involve both complementarity and interference rather than uniformly improving performance. Third, it uncovers a counterintuitive finding that multimodal inputs paired with unimodal annotations can sometimes outperform fully multimodal supervision. Together, these results provide new empirical insights into cross-modal dynamics and inform the design of multimodal emotion recognition systems.

7. Limitations

Since the aim of this paper is to investigate cross-modal interactions, all experiments share the same hyperparameters as the multimodal setup (M-TAV) to ensure a fair comparison across different configurations. This design choice, however, may lead to suboptimal performance in single-modality setting, for example, the text-only baseline predicted all instances as *neutral*, as shown in Figure 2.

This study uses only a single dataset, as, to the best of our knowledge, it is the only publicly available resource containing both independent unimodal and multimodal categorical emotion labels suitable for the proposed experimental design.

Although the research question itself is not entirely novel, this work represents a systematic empirical study to validate a widely assumed hypothesis regarding cross-modal influences on unimodal emotion recognition.

The Tensor Fusion Network (TFN) was chosen because it is sufficiently expressive and efficient for exploring multimodal emotion interactions, while avoiding mechanisms such as attention that explicitly promote or suppress information. This ensures that all modality information is preserved in the model. More recent advanced models, which incorporate selective fusion mechanisms, may not be suitable for this purpose.

Large language models are not employed in this study, primarily because the audio and visual data are private and raise ethical concerns regarding privacy and consent. Only pre-extracted feature packages are available and used for this study.

8. Acknowledgements

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (174P07826). We would also like to thank the anonymous reviewers for their valuable and constructive feedback.

9. Bibliographical References

Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. Context in emotion perception. *Current directions in psychological science*, 20(5):286–290.

Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211.

Carlos Busso, Srinivas Parthasarathy, Alec Burmania, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE Transactions on Affective Computing*, 8(1):67–80.

Jacob 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Quanqi Du, Loic De Langhe, Els Lefever, and Veronique Hoste. 2025a. Ldw : Label divergence weighting for multimodal sentiment analysis. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, pages 1–10. Association for Computing Machinery.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2023. Unimodalities count as perspectives in multimodal emotion annotation. In *2nd Workshop on Perspectivist Approaches to NLP (NLPerspectives 2023), co-located with the 26th European Conference on Artificial Intelligence (ECAI 2023)*, volume 3494. CEUR-WS. org.

Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025b. UniC: A dataset for emotion analysis of videos with multimodal and unimodal labels. *Language Resources and Evaluation*, 59:2857–2892.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: The munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Kaouther Ezzameli and Hela Mahersia. 2023. Emotion recognition from unimodal to multimodal analysis: A review. *Information Fusion*, 99:101847.

Yanping Fu, Zhiyuan Zhang, Ruidi Yang, and Cuiyou Yao. 2024. Hybrid cross-modal interaction learning for multimodal sentiment analysis. *Neurocomputing*, 571:127201.

Samira Hazmoune and Fateh Bougamouza. 2024. Using transformers for multimodal emotion recognition: Taxonomies and state of the art review. *Engineering Applications of Artificial Intelligence*, 133:108339.

Ma Junchi, Hassan Nazeer Chaudhry, Farzana Kulsoom, Yang Guihua, Sajid Ullah Khan, Sujit Biswas, Zahid Ullah Khan, and Faheem Khan. 2025. Multicausenet temporal attention for multimodal emotion cause pair extraction. *Scientific Reports*, 15(1):19372.

Mustaqeem Khan, Phuong-Nam Tran, Nhat Truong Pham, Abdulmotaleb El Saddik, and Alice Othmani. 2025. Memocmt: multimodal emotion recognition using cross-modal transformer-based feature fusion. *Scientific reports*, 15(1):5473.

Ayush Kumar and Jithendra Vepa. 2020. Gated mechanism for attention based multi modal sentiment analysis. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.

Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and CKM Lee. 2023. Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing. *Computer science review*, 48:100545.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. Make Acoustic and Visual Cues Matter: CH-SIMS v2.0 Dataset and AV-Mixup Consistent Module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 247–258, New York, NY, USA. Association for Computing Machinery.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. Mediapipe: A framework for building perception pipelines.

Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2010. Decision level combination of multiple modalities for recognition and analysis of emotional expression. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2462–2465. IEEE.

Bei Pan, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. 2023. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. *Neurocomputing*, 561:126866.

Rosalind W. Picard. 1997. *Affective Computing*. MIT press, Cambridge.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. Dynamic graph neural ODE network for multimodal emotion recognition in conversation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 256–268, Abu Dhabi, UAE. Association for Computational Linguistics.

Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. 2023. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8967–8989, Singapore. Association for Computational Linguistics.

Sanja Stajner and Roman Klinger. 2023. Emotion analysis from texts. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 7–12, Dubrovnik, Croatia. Association for Computational Linguistics.

Xin Sun, Xiangyu Ren, and Xiaohao Xie. 2024. A novel multimodal sentiment analysis model based on gated fusion and multi-task learning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8336–8340.

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

Harald G Wallbott and Klaus R Scherer. 1986. Cues and channels in emotion recognition. *Journal of personality and social psychology*, 51(4):690.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2023. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921.

Matthias J Wieser and Tobias Brosch. 2012. Faces in context: A review and systematization of contextual influences on affective face processing. *Frontiers in psychology*, 3:471.

Yunhe Xie, Chengjie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. 2025. A dual contrastive learning framework for enhanced multimodal conversational emotion recognition. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4055–4065, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

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

Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2024. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects. *Expert Systems with Applications*, 237:121692.