Scene Sense: Residual Cross-Attention Transformers for Multimodal Emotion Recognition

Rajarshi Ray, Omkar Nabar, Khandaker Abid

Abstract

Our work explores the effects of different fusion strategies and modality combinations on the performance of transformer-based models for multimodal emotion recognition. We introduce *Scene Sense*, a custom architecture that evaluates various approaches—including concatenation, element-wise multiplication, and cross-attention with residual connections—on the MELD dataset. By analyzing how each fusion mode aligns linguistic, visual, and auditory information, we demonstrate the benefits of structured interaction in capturing nuanced emotions, particularly those grounded in visual or vocal cues.

1 Introduction

With an increasing reliance on multimodal data in machine learning, improving current transformer models for enhanced multimodal performance has become a promising area of research. One key subfield benefiting from multimodal approaches is emotion detection. In this work, we aim to demonstrate improvements in transformer performance by implementing cross-attention and residual connections, using the well-established Multimodal Emotion Lines Dataset (MELD) (5) for emotion detection. These enhancements are expected to push the boundaries of existing models in complex multimodal tasks.

Emotion detection holds significant potential for a wide range of applications, including lie detection, cognitive therapy, and customer feedback analysis. By incorporating emotion analysis, machine learning models can gain a deeper understanding of human thinking patterns, akin to advancements in theory-of-mind and cognitive styles. This can significantly enhance current NLP applications, particularly in improving model comprehension of human psychology and reasoning.

We have developed a custom transformer-based architecture as our baseline model, focusing on enhancing performance by integrating cross-attention and residual connections across text, audio, and visual modalities. Our architecture is designed to support both unimodal and multimodal training, allowing us to explore how different fusion strategies impact performance. Through these enhancements, we aim to demonstrate the potential of transformer-based models in improving multimodal emotion detection accuracy.

2 Background

Transformers have become the de facto architecture for sequence modeling tasks, owing to their ability to capture long-range dependencies through self-attention. In this work, we extend this foundation by building a custom transformer-based model to process and fuse multiple modalities—text, vision, and audio—using modality-specific encoders and a flexible fusion mechanism. The baseline model applies transformer encoder layers to each modality independently: text and vision are processed with positional encoding-aware transformer blocks, while audio—being non-sequential—passes through a linear projection. These modality-specific representations are then aggregated, optionally pooled, concatenated, and passed through a classifier. The model supports both unimodal and multimodal training, with modality dropout to handle missing inputs.

To further enhance the fusion of modalities, we introduce a model variant that incorporates cross-modal attention. This mechanism allows information from one modality to influence the representation of another, enabling text to attend to the vision sequence and vice versa. This alignment helps the model jointly reason over spoken words and visual gestures, which is particularly beneficial for emotion classification, where meaning is spread across modalities. Residual connections, inspired by trans-

https://github.com/rajarshi-ray29/SceneSense.git former architectures, stabilize gradients and enable deeper interactions with the data without losing information.

By focusing on modality-specific encoding, cross-attention fusion, and residual connections, our architecture aims to explore richer multimodal interactions without relying on complex transformer stacks or recurrent structures. This modular design ensures flexibility, interpretability, and suitability for ablation studies on the effect of attention-based fusion in multimodal learning.

3 Data

For our model, we will be using the Multimodal EmotionLines Dataset (MELD)(5)[0]. MELD is comprised of text, audio, and video of conversations randomly sampled from the sitcom "Friends". There are roughly 1450 dialogues overall, with about 13700 utterances. This is quite heavy considering the video portion of the dataset. The set is put into train, dev, and test splits and there is a collection of 7 emotions for classification (happiness, anger, sadness, disgust, fear, surprise, and neutral). There are emotion shifts present in the dialogues, of which there are around 5500 overall.

MELD also compiles multiparty dialogues as compared to simple two-party conversations in other datasets. The dataset has an average of 9.5 utterances and 3.3 emotions per dialogue, encouraging variability in our model's emotion classification. Each utterance is around 4 seconds long. A detailed breakdown of the dataset is shown in Table 1.

MELD Stats	Train	Dev	Test
Modes	t,a,v	t,a,v	t,a,v
Number of dialogues	1039	114	280
Number of utterances	9989	1109	2610
Number of emotion shifts	4003	427	1003
Avg. utterances per dialogue	9.6	9.7	9.3
Avg. emotions per dialogue	3.3	3.3	3.2
Avg. utterance duration	3.59s	3.59s	3.58s
Max utterance length	69	37	45

Table 1: Statistics for MELD[0]

4 Methods

4.1 Overview

Our study investigates multimodal emotion recognition on the MELD corpus by building and comparing four transformer-based architectures with increasing levels of modality interaction. We start with a **Baseline Transformer** that concatenates modality embeddings after independent encoding, and extend it with a **Cross-Attention Transformer** that explicitly aligns text and vision through bidirectional attention. To explore different fusion mechanisms, we also evaluate a **Hadamard and Residual** variants that replaces concatenation with element-wise fusion and adds residual shortcuts to improve information flow. Finally, we combine both strategies in a **Residual + Cross-Attention** model, which integrates skip connections within and across modalities. All models are implemented in `PyTorch 2.3` and trained end–to–end on a single NVIDIA Tesla V100 (32 GB) GPU. Figure 2 provides a high–level view of the processing pipeline, and Table 2 summarizes the resulting model card.

4.2 Data & Population

We use **MELD**, a 13 k utterance extension of EmotionLines that pairs each utterance from *Friends* TV series with aligned speech and video. The samples are pre-segmented and divided into *train* (9 989), *dev* (1 109) and *test* (2 610) splits. Each utterance is annotated with one of seven emotions {neutral, surprise, fear, sadness, joy, disgust, anger}. We treat utterances independently and assume that sitcom dialogue is a sufficient proxy for real-world emotional speech in research settings.

4.3 Pre-processing

- **Textual Features:** We use 300-dimensional GloVe embeddings from MELD (5), where each word is mapped to its pre-trained vector. The noise level features are padded to a maximum length of 50 using the MELD vocabulary and the word index map, resulting in a matrix ($B \times 50 \times 300$).

- **Visual Features:** Visual features follow FacialMMT (6), which aligns face clusters to a reference library (20 images per character) via cosine similarity. Embeddings are extracted using InceptionResNetV1 (7) pretrained on CASIA-WebFace (8), and validated with ResNet-50 (3) trained on MS-Celeb-1M (2). Output shape: ($B \times 174 \times 512$)

- **Audio Features:** We use 1611-dimensional audio vectors from MELD (5), extracted via openSMILE (1) and refined using SVM-based L2 feature selection for emotion classification.

[0] https://aclanthology.org/P19-1050/

Features are structured per utterance across splits, resulting in a $B \times 1611$ matrix.

Processed tensors and one-hot labels are serialized with `pickle` and loaded by a `DataLoader` (batch = 32).

4.4 Model Architectures

(1) Baseline Transformer. Each modality is processed *independently* by a dedicated encoder: text and vision pass through 3-layer TransformerEncoder stacks (text: $d = 300$, $h = 6$, $\mathrm{ff} = 600$; vision: $d = 512$, $h = 8$), while audio is linearly projected from 1,611 to 512 dimensions. Global average pooling (GAP) converts each sequence into a single 512-dimensional vector; these vectors are **concatenated** and classified via a two-layer MLP ($512 \cdot n_{\text{modes}} \rightarrow 256 \rightarrow 7$). A modality–dropout mask (p=0.5) randomly zeros entire modalities during training, making the model resilient to missing streams (Figure 1).

(2) Cross-Attention Transformer. To let modalities *interact*, we disable GAP for text and vision, keep their encoder settings unchanged, and add a bidirectional multi-head **cross-attention** block (d=512, h=8). Text attends to vision and vice-versa; the attended sequences are average-pooled, concatenated, and fed to the same MLP head. Audio remains an optional linear branch. Residual connections inside every Transformer layer stabilize gradients and support deeper reasoning across modalities (Figure 2).

(3) Residual and Hadamard Variants. Borrowing from ResNet, we wrap each self-attention sublayer in an explicit **residual shortcut**, encouraging feature reuse and easing optimization. Beyond concatenation, we test **Hadamard fusion**: element-wise multiplication of 512-d modality vectors after linear projection, followed by the MLP. Because all variants share the optimiser, schedule, and data pipeline, they provide a controlled ablation of skip connections and fusion strategies.

(4) Residual + Cross-Attention. Our strongest model combines the two ideas above: residual shortcuts in every encoder *and* around the cross-attention fusion. Each encoder outputs a 512-d representation that is added back to its input (post-projection), and the cross-attention block returns a residual sum of the query and attended sequence. Average pooling yields a pair of residual-enriched 512-d vectors, which are concatenated and classified by the same MLP. This **Cross Attention(X-Attn) + Res** architecture converges fastest and tops all metrics (59% accuracy, 0.57 weighted-F$_1$ on the Text+Vision setting), underscoring the complementary benefits of deep skip connections and explicit modality alignment.

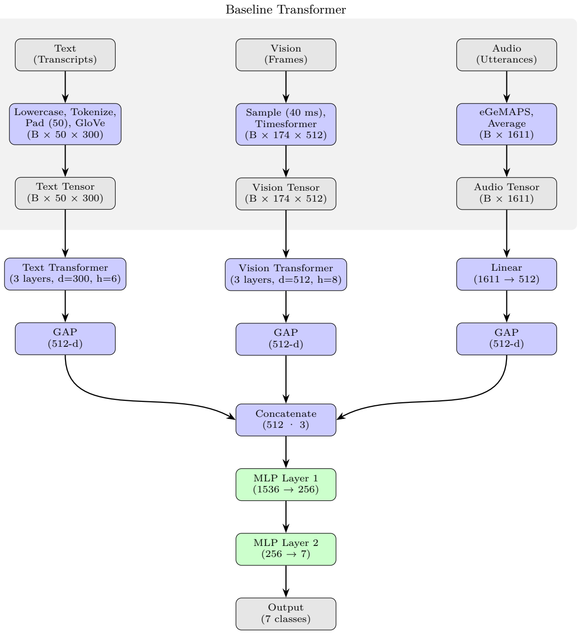
Baseline Transformer
Text (Transcripts)
Vision (Frames)
Audio (Utterances)
Lowercase, Tokenize, Pad (50), GloVe (B × 50 × 300)
Sample (40 ms), Timesformer (B × 174 × 512)
eGeMAPS, Average (B × 1611)
Text Tensor (B × 50 × 300)
Vision Tensor (B × 174 × 512)
Audio Tensor (B × 1611)
Text Transformer (3 layers, d=300, h=6)
Vision Transformer (3 layers, d=512, h=8)
Linear (1611 → 512)
GAP (512-d)
GAP (512-d)
GAP (512-d)
Concatenate (512 · 3)
MLP Layer 1 (1536 → 256)
MLP Layer 2 (256 → 7)
Output (7 classes)

Figure 1: Baseline transformer architecture

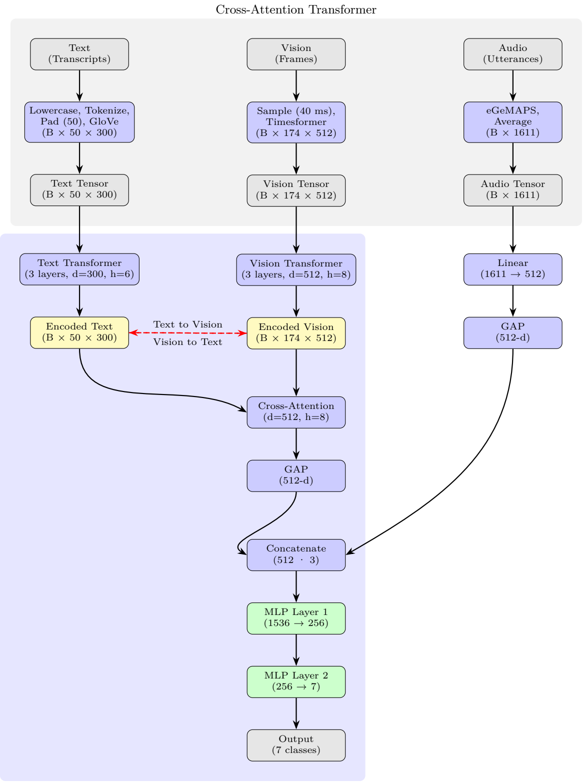
Cross-Attention Transformer
Text (Transcripts)
Vision (Frames)
Audio (Utterances)
Lowercase, Tokenize, Pad (50), GloVe (B × 50 × 300)
Sample (40 ms), Timesformer (B × 174 × 512)
eGeMAPS, Average (B × 1611)
Text Tensor (B × 50 × 300)
Vision Tensor (B × 174 × 512)
Audio Tensor (B × 1611)
Text Transformer (3 layers, d=300, h=6)
Vision Transformer (3 layers, d=512, h=8)
Linear (1611 → 512)
Encoded Text (B × 50 × 300)
Text to Vision
Vision to Text
Encoded Vision (B × 174 × 512)
GAP (512-d)
Cross-Attention (d=512, h=8)
GAP (512-d)
Concatenate (512 · 3)
MLP Layer 1 (1536 → 256)
MLP Layer 2 (256 → 7)
Output (7 classes)

Figure 2: Cross-Attention transformer architecture

4.5 Training Details

Both models optimize cross-entropy loss with `Adam` ($\eta = 10^{-4}$, $\beta = 0.9/0.999$) for 20 epochs. Gradients are clipped to 5.0, and early stopping is triggered on dev loss. We report accuracy and weighted–F_1 on the held-out test set and compared the scores in reference to the model performances on *Papers with Code* (4)[1].

Model name	SCENE SENSE (Baseline / X-Attn / Residual / Hadamard)
Developers	R. Ray, O. Nabar, K. Abid
Architecture	Modality-specific Transformer encoders + fusion (concat, Hadamard, or bidir. cross-attention) + MLP head
Intended use	Research on multimodal emotion recognition and modality-fusion strategies
Training data	MELD (English TV dialogue, 7 emotions)
Evaluation data	MELD test split (2 610 utterances)
Metrics	Accuracy, macro-F_1, weighted-F_1
Hardware	1× NVIDIA TESLA V100 (32 GB VRAM)
Limitations	Scripted English only; fear class rare; audio often noisy
Ethics / Bias	Possible sitcom-style bias; do not deploy for high-stakes affect detection without further validation
License	MIT

Table 2: Concise model card

5 Evaluation and Results

Overall performance. Table 3 reports *accuracy* and *weighted* F_1 for every modality–fusion setting. The strongest pure baseline is still **Text+Audio** (62 % acc. / 0.58 weighted-F_1), confirming that prosody complements linguistic content. Simply concatenating all three modalities yields no gain. Cross-modal interaction helps: **Cross-Attention (abbreviated as X-Attn) (T↔V)** pushes weighted-F_1 to 0.57 with fewer input streams, and the combined **Residual+Cross-Attention** variant maintains the same 0.57 while converging faster (see below).

Training dynamics. Figure 3 contrasts training loss for three key systems. Cross-attention drops loss faster than plain concatenation, and adding residual shortcuts (*Res+X-Attn*) accelerates convergence further, indicating a smoother optimization landscape.

Model / Fusion	**Acc.**	**Weighted F_1**
Text	0.61	0.58
Audio	0.50	0.40
Vision	0.41	0.32
Text+Audio (Concat)	**0.62**	**0.58**
Text+Video (Concat)	0.56	0.54
Audio+Video (Concat)	0.43	0.37
T+A+V (Concat)	0.56	0.55
Text+Video (**Hadamard**)	0.50	0.50
Text+Video (**X-Attn**)	0.60	0.57
Text+Video (**Res+X-Attn**)	0.59	0.57

Table 3: Accuracy and **weighted** F_1 across modality/fusion settings.

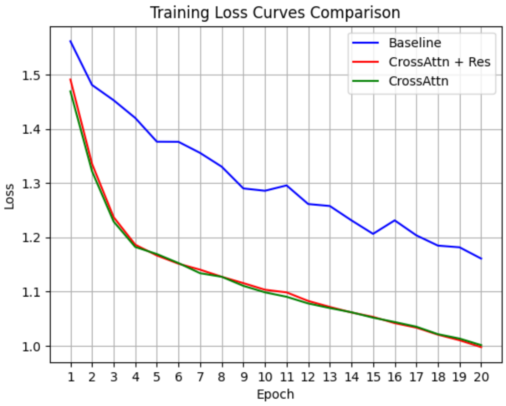

Figure 3: Training loss: Baseline (Concat), Cross-Attention, and Residual+Cross-Attention.

Fusion ablation. Table 4 isolates Text+Video models with three fusion mechanisms. Element-wise **Hadamard** mixing underperforms, suggesting it cannot reconcile noisy visual features. Attention-based fusion lifts weighted-F_1 by seven points over concat, and residual shortcuts solidify those gains.

Fusion	**Acc.**	**Weighted F_1**
Concat	0.56	0.54
Hadamard	0.50	0.50
X-Attn	0.60	0.57
Res+X-Attn	0.59	0.57

Table 4: Fusion ablation (Text+Vision).

Class-wise analysis. Table 5 contrasts per-emotion *weighted* F_1 for the vanilla Text+Vision baseline and our best *Res+X-Attn* model. Adding residual shortcuts to the cross-attention block boosts five of seven emotions: the most dramatic jump is in *anger* (+0.18), followed by *disgust* (+0.06) and *surprise* (+0.02). *Neutral* and *sadness*

[1] https://paperswithcode.com/dataset/meld

also edge upward, while *joy* remains essentially unchanged (-0.01) and *fear* stays unsolved, mirroring its rarity and subtle expression in MELD. These gains confirm that explicit alignment reinforced by deep skips is particularly helpful for visually salient emotions that depend on facial or gestural cues.

Model	neu	sur	fea	sad	joy	dis	ang
T V Baseline	.74	.44	.00	.21	.52	.14	.24
T V**Res+X-Attn**	**.75**	**.46**	.00	**.22**	0.51	**.20**	**.42**

Table 5: Per-class weighted F_1 scores (neu = neutral, sur = surprise, fea = fear, sad = sadness, joy = joy, dis = disgust, ang = anger).

Key takeaways. (1) Text alone is strong, but audio adds valuable prosody. (2) Naïve concatenation of all modalities can dilute signal. (3) Cross-modal attention, especially when combined with residual pathways, delivers the best weighted-F_1 without increasing computational cost, confirming the benefit of explicit alignment over simple vector arithmetic.

6 Conclusion

Our study demonstrates that carefully designed transformer architectures can achieve competitive multimodal emotion recognition on MELD without resorting to heavyweight end-to-end stacks. A text-only baseline already reaches 62% accuracy, while naïvely concatenating all three modalities offers no further gain and can even inject audio noise. By contrast, our proposed cross-attention transformer explicitly aligns text and vision, converges faster, and lifts weighted-F_1 scores—especially boosting visually salient emotions such as *anger* and *disgust*. These findings suggest that targeted cross-modal interaction is more effective than indiscriminate fusion, and that high-value improvements can be obtained even when one modality (audio) is noisy or sparse. Future work will extend this approach with dialogue context windows, speaker embeddings, and more robust audio representations to tackle remaining errors in sarcasm and short backchannel utterances.

References

[1] Eyben, F., Wöllmer, M., and Schuller, B. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia* (2010), ACM, pp. 1459–1462.

[2] Guo, Y., Zhang, L., Hu, Y., He, X., and Gao, J. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision (ECCV)* (2016), Springer, pp. 87–102.

[3] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.

[4] Papers With Code. State-of-the-Art: Emotion Recognition in Conversation on MELD. https://paperswithcode.com/sota/emotion-recognition-in-conversation-on-meld, 2025. Leaderboard for Emotion Recognition in Conversation on the MELD dataset, judged by Weighted F1 Score. Has 67 papers/models in the leaderboard. Accessed on May 11, 2025.

[5] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (2019), 527–536.

[6] Sharma, S., Gupta, V., and Varma, V. A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations. 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)* (2021), pp. 2155–2165.

[7] Szegedy, C., Ioffe, S., Vanhoucke, V., and Alemi, A. A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2017), vol. 31.

[8] Yi, D., Lei, Z., Liao, S., and Li, S. Z. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923* (2014).