Integrating Acoustic Information with Semantics for Interpretable Speech Emotion Recognition

Moritz Lenhard

Chair of Health Informatics
School of Computation, Information and Technology (Informatik)
Technical University of Munich

Integrating Acoustic Information with Semantics for Interpretable Speech Emotion Recognition

Moritz Lenhard

Thesis for the attainment of the academic degree

Bachelor of Science (B.Sc.)

at the School of Computation, Information and Technology (Informatik) of the Technical University of Munich.

Examiner:
Prof. Dr.-Ing. habil. Björn Schuller

Supervisor:
Andreas Triantafyllopoulos

Submitted:
Munich, 15.02.2026

I hereby declare that this thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

Munich, 15.02.2026 Moritz Lenhard

AI Disclosure

Generative AI was utilized as a tool for linguistic drafting, refinement, brainstorming, the implementation of specific code components, and the generation of schematic visualizations. However, the research direction, experimental methodology, and critical evaluation of all findings were conducted independently, and I remain solely responsible for the final content.

Abstract

Effective Speech Emotion Recognition (SER) requires the integration of paralinguistic acoustics with linguistic semantics. However, current multimodal approaches often rely on "black-box" latent fusion, which obscures decision boundaries and struggles to resolve conflicting cues, such as distinguishing high-arousal *Happiness* from low-arousal *Neutrality* when semantic content is ambiguous. This thesis investigates the trade-offs between latent architectural complexity and explicit domain modeling to bridge this "Modality Gap."

We systematically evaluate two fusion paradigms on the acted IEMOCAP dataset: (1) a *Hierarchical Cross-Modal Transformer* (HCMT) evaluating latent interaction under fixed temporal granularity, and (2) a proposed *Arousal-Augmented Fusion* framework that explicitly injects predicted psychometric arousal into the classification decision boundary. We utilize Wav2Vec 2.0 and BERT as feature extractors to capture acoustic and semantic information, respectively.

Our experiments demonstrate that architectural complexity does not guarantee superior performance in data-constrained environments. The *Arousal-Augmented Fusion* model achieves an Unweighted Average Recall (UAR) of **0.743** on the acted data, outperforming the baseline Early Fusion (0.726) and marginally surpassing the HCMT (0.737). Error analysis reveals that while Transformers suffer from "recency bias" due to limited training data, explicit psychometric injection acts as an effective gating mechanism, correcting 25% of *Happy*-to-*Neutral* misclassifications found in the baseline. These findings suggest that for SER on acted corpora, incorporating explicit psychometric priors offers a more stable and interpretable alternative to complex latent attention mechanisms.

Contents

List of Abbreviations

1 Introduction

1.1 Context

Speech Emotion Recognition (SER) constitutes a pivotal component of Affective Computing, aiming to enable Human-Computer Interaction (HCI) systems to discern emotional states from vocal signals. Applications of SER range from automated call center analysis to objective assessment in mental health monitoring [1]. While early approaches relied primarily on acoustic signal processing, recent literature emphasizes that human communication is inherently multimodal, encoding affective information simultaneously through linguistic content (semantics) and paralinguistic cues (acoustics) [2].

The integration of these modalities addresses a fundamental limitation in unimodal modeling known as the “Valence-Arousal Gap.” Acoustic features, such as energy and pitch dynamics, are robust indicators of *Arousal* (emotional intensity) but often fail to distinguish between emotions of similar intensity but opposite polarity, such as *Anger* and *Happiness* [3]. Conversely, semantic features derived from text transcripts excel at capturing *Valence* (positivity or negativity) but may miss the emotional context provided by tone, particularly in cases of sarcasm or irony. Consequently, state-of-the-art systems increasingly adopt multimodal fusion strategies to synthesize these complementary information streams [4, 5].

Technologically, the field has transitioned from hand-crafted low-level descriptors (e.g., extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) [6]) to representations learned by self-supervised foundation models. Architectures such as Wav2Vec 2.0 [7] and BERT [8] allow for the extraction of rich, contextualized embeddings from raw audio and text, respectively. However, the fusion of these high-dimensional embeddings often results in opaque, “black-box” architectures. This lack of interpretability poses a significant barrier to the adoption of SER systems in safety-critical domains, necessitating architectures that not only classify emotions accurately but also explicate the contribution of each modality to the decision process.

1.2 Problem Statement

While the efficacy of SER systems has improved significantly with the advent of deep learning, three critical challenges remain unresolved in the transition from unimodal to multimodal architectures.

1. The Modality Gap and Semantic Ambiguity Speech is inherently multimodal, consisting of linguistic content (semantics) and paralinguistic acoustic cues (prosody). Unimodal models are fundamentally limited: acoustic models often struggle to distinguish states with similar arousal levels but differing valence (e.g., *Fear* vs. *Anger*), while semantic models fail when emotional intent contradicts the literal meaning of words, as in sarcasm or irony [2]. Consequently, relying on a single modality introduces a *modality gap*, where the absence of complementary information leads to irreducible classification error. The challenge lies not merely in combining these modalities, but in resolving cases where they provide conflicting signals.

2. The "Black Box" Nature of Latent Fusion Current state-of-the-art approaches typically employ "late fusion" or complex cross-modal attention mechanisms (e.g., Multimodal Transformers) to integrate acoustic and text embeddings [9, 4]. While these methods often yield high accuracy, they operate as "black boxes." The fusion occurs in a high-dimensional latent space, obscuring the decision boundary. It remains opaque whether a classification of *Anger* is driven by aggressive semantic tokens or high acoustic energy. This

lack of **intrinsic interpretability** is a barrier to clinical and safety-critical applications, where understanding the *why* behind a prediction is as critical as the prediction itself.

3. Architectural Complexity vs. Data Scarcity Recent literature prioritizes complex architectures, such as the Cross-Modal Gated Interaction [5], which attempt to learn alignment between audio and text from scratch. However, common SER benchmarks like Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) are relatively small (approx. 10,000 utterances) compared to the massive corpora used in Natural Language Processing (NLP) [10]. There is a risk that complex attention mechanisms fail to converge or exhibit "attention collapse" on such limited data, effectively ignoring one modality in favor of the other [11]. It remains an open research question whether explicit, domain-grounded fusion strategies (injecting psychometric priors like Arousal/Valence) might offer a more stable and interpretable alternative to implicit latent attention for datasets of this scale.

1.3 Research Questions

To address the challenges of modality fusion, semantic ambiguity, and black-box opacity in SER, this thesis investigates the trade-offs between architectural complexity, temporal granularity, and the use of explicit domain knowledge. We formulate the following three research questions to guide our experimental design:

RQ1: Temporal Granularity vs. Global Aggregation.
Does strict word-level temporal alignment between acoustic and semantic modalities yield superior performance compared to global utterance-level aggregation?
Standard multimodal approaches often assume that fine-grained alignment (e.g., matching specific acoustic frames to semantic tokens via Long Short-Term Memory (LSTM) networks or Transformers) is necessary to resolve ambiguity. This question investigates whether the computational cost of forced alignment is justified for short, acted utterances in IEMOCAP, or if global statistical aggregation (e.g., mean pooling) provides a sufficiently robust representation.

RQ2: Explicit Psychometric Injection.
Can the "modality gap" be bridged more effectively by explicitly injecting predicted psychometric primitives (Arousal/Valence) into the fusion layer?
We hypothesize that the specific confusion between active and passive emotions (e.g., *Happiness* vs. *Neutral*) stems from a lack of explicit intensity modeling in standard concatenation fusion. This question tests whether injecting a disentangled scalar signal from a dedicated regressor acts as an effective "soft gate" to resolve these ambiguities.

RQ3: Latent Expressivity vs. Inductive Bias.
Does implicit latent alignment (via Cross-Modal Attention) offer superior interpretability or performance compared to explicit psychometric injection in data-constrained environments?
While Hierarchical Cross-Modal Transformer (HCMT) represent the state-of-the-art in large-scale multimodal learning, they rely on learning alignment from scratch. This question critically evaluates whether these complex attention mechanisms converge effectively on the limited-size IEMOCAP dataset, or if they suffer from "attention collapse" compared to the stronger inductive bias provided by the explicit domain priors proposed in RQ2.

1.4 Contributions

This thesis critically examines the architectural trade-offs between latent representation learning and explicit domain modeling in multimodal SER. By systematically evaluating fusion strategies on the IEMOCAP dataset, we challenge the prevailing assumption that increased architectural complexity (e.g., cross-modal

transformers) inherently yields superior performance in data-constrained environments. The specific contributions are as follows:

1. **Development of the Arousal-Augmented Fusion Framework:** We propose and validate a modular fusion architecture that explicitly injects predicted psychometric primitives (Arousal and Valence) into the classification decision boundary. Unlike standard late fusion or "black-box" concatenation, this method mathematically exposes the interaction between acoustic intensity and semantic content. We demonstrate that this explicit injection acts as a soft gating mechanism, specifically resolving the “modality gap” between high-energy emotions (e.g., *Happiness*) and neutral states. This approach achieves a Unweighted Average Recall (UAR) of **0.743** on IEMOCAP (Session 5), statistically outperforming both the baseline early fusion (0.726) and complex transformer variants.

2. **Mechanistic Analysis of Dimensional Interaction:** Beyond aggregate performance metrics, we provide a granular analysis of *how* psychometric primitives influence the multimodal decision boundary. We demonstrate that Arousal injection functions as a discriminative **energy gate**, effectively resolving the ambiguity between high-activation states (e.g., *Happiness*) and *Neutrality*. Conversely, we identify that Valence injection acts as a **negative exclusion filter**; despite suffering from regressor calibration issues (scale compression), it successfully constrains the search space for passive negative emotions (e.g., *Sadness*) by suppressing positive class probabilities. This analysis provides an interpretable explanation for the performance gains observed in the Arousal-Augmented framework.

2 Related Work

2.1 Unimodal Representations

Effective multimodal emotion recognition relies fundamentally on the quality of the feature representations extracted from the constituent modalities. This section traces the evolution of acoustic and semantic feature extractors from hand-crafted descriptors to the current paradigm of self-supervised foundation models.

2.1.1 Acoustic Representations

Acoustic feature extraction has historically aimed to identify paralinguistic cues that correlate with emotional dimensions, particularly Arousal (e.g. pitch, energy, voice quality).

Hand-Crafted Features Early approaches relied on expert-defined feature sets. The eGeMAPS became a standard in the field, reducing high-dimensional audio signals into a compact vector of 88 functionals (e.g., mean, standard deviation) applied to Low-Level Descriptors (LLDs) like frequency, jitter, and shimmer [6]. While these features offer domain interpretability by explicitly mapping physical properties to scalar values, they often fail to capture complex temporal dynamics and subtle linguistic nuances compared to deep representations [2].

Self-Supervised Learning (SSL) The state-of-the-art has shifted toward SSL models trained on massive unlabeled corpora.

- **Wav2Vec 2.0:** Baevski et al. proposed learning speech representations directly from the raw waveform using a contrastive task over quantized latent representations [7]. This model captures both acoustic and coarse linguistic information, making it a standard feature extractor for SER.
- **HuBERT:** Concurrently, Hsu et al. introduced Hidden Unit BERT (HuBERT), which predicts cluster assignments of masked audio segments [12]. Wagner et al. demonstrated that fine-tuning the transformer layers of these foundation models is critical for improving Valence prediction, traditionally a weak point for acoustic-only models [13].
- **Weakly-Supervised ASR Models (Whisper):** More recently, Radford et al. introduced Whisper, a model trained on 680,000 hours of multilingual data using weak supervision for Automatic Speech Recognition (ASR) [14]. Unlike Wav2Vec 2.0, which optimizes a contrastive objective, Whisper is optimized for transcription accuracy. While highly robust to noise, recent benchmarks suggest that for paralinguistic tasks like SER, SSL models (like Wav2Vec 2.0) often yield more discriminative embeddings than ASR-focused models, as the latter may suppress prosodic variations irrelevant to text generation [15].

2.1.2 Semantic Representations

Semantic analysis in SER aims to resolve the “Valence Gap,” as acoustic intensity often correlates well with Arousal but poorly with emotional polarity (e.g., distinguishing *Happy* from *Angry*).

Contextual Embeddings (BERT) The introduction of the Transformer architecture revolutionized semantic modeling. We utilize Bidirectional Encoder Representations from Transformers (BERT) [8], which solves the context limitation of static embeddings (e.g., Word2Vec [16]) by training on a Masked Language Model (MLM) objective. BERT considers bidirectional context simultaneously, producing dynamic embeddings where a token's representation shifts based on its neighbors [17]. This allows for the detection of sentiment-shifting cues (e.g., "not bad" vs. "bad") essential for Valence detection.

Generative Large Language Models A recent trend in affective computing involves prompting Generative Large Language Models (LLMs) (e.g., GPT-4, Llama) to perform zero-shot or few-shot emotion classification [1]. While these models demonstrate emergent reasoning capabilities suitable for complex mental health tasks, they typically operate as "black boxes" via text generation APIs. In this thesis, we prioritize distinct encoder-based architectures (BERT) over generative LLMs to maintain precise control over the fusion mechanism (Section 2.2) and to enable the explicit mathematical modelling of the Arousal-Valence interaction, which is obscured in the latent states of massive generative models.

2.2 Multimodal Fusion in SER

The integration of acoustic and semantic information is the central algorithmic challenge in multimodal SER. Fusion strategies must address two fundamental difficulties: the heterogeneity of the data (continuous audio signals vs. discrete text tokens) and the asynchronous nature of human expression, where an emotional cue in the voice may precede or follow the corresponding keyword. This section categorizes existing approaches into classical fusion paradigms, transformer-based cross-modal attention, and recent hierarchical interaction models.

2.2.1 Classical Paradigms: Early and Late Fusion

Historically, multimodal fusion has been categorized by the stage at which integration occurs. **Early Fusion** (or feature-level fusion) involves concatenating feature vectors from different modalities into a single representation vector before feeding it into a classifier. While simple to implement, this approach assumes a strict temporal synchronicity between modalities that is rarely present in natural speech [4]. Furthermore, simple concatenation can lead to the "dominance" of one modality if the feature spaces are not properly normalized or balanced.

Conversely, **Late Fusion** (or decision-level fusion) trains independent unimodal classifiers and aggregates their predictions using voting schemes, averaging, or a meta-learner [5]. While this preserves the specific dynamics of each modality, it fails to capture low-level cross-modal interactions, such as the modification of a word's meaning by a sarcastic tone. Consequently, simple late fusion is often insufficient because the decision boundaries are formed in isolation.

2.2.2 The Transformer Era: Cross-Modal Attention and Gating

The advent of the Transformer architecture [17] introduced mechanisms to model dependencies between asynchronous sequences without the need for rigid temporal alignment. Two foundational architectures in this domain form the basis for the latent interaction models investigated in this thesis.

Cross-Modal Attention (MulT) Tsai et al. introduced the Multimodal Transformer (MulT), which shifted the paradigm from alignment to attention [18]. Instead of forcing alignment between audio and text, MulT employs directional pairwise cross-modal attention. In this mechanism, the features of one modality (e.g., text) serve as the *Query* (Q), while the features of another (e.g., audio) serve as *Key* (K) and *Value* (V). This allows the model to latently adapt the semantic stream by attending to relevant acoustic events, regardless of their temporal distance.

Multimodal Adaptation Gate (MAG) Rahman et al. proposed an alternative to pure attention: the Multimodal Adaptation Gate (MAG) [9]. Rather than a full cross-attention mechanism, MAG injects a shift vector derived from non-verbal modalities into the internal representations of large pre-trained language models like BERT. This "gating" approach treats acoustics as a context signal that modulates the semantic embedding space. This concept underpins our rationale for the *Arousal-Augmented Fusion* (Section 3.4), where we inject explicit psychometric scalars to modulate the decision boundary, rather than implicit latent vectors.

2.2.3 State-of-the-Art Interaction Models

Recent literature has pushed towards increasingly complex architectures to maximize performance, often focusing on robustness against ASR errors or missing modalities.

Recent advances have moved beyond simple fusion towards mechanisms that explicitly control information flow. Gao et al. represent the state-of-the-art in *explicit information extraction*, utilizing a Cross-Modal Gated Interaction (CmGI) module to explicitly filter semantic features based on ASR error predictions [5]. While this validates the move away from black-box fusion, their approach focuses primarily on robustness against linguistic noise. In contrast, this thesis employs explicit injection to model *psychometric dimensions* (Arousal and Valence), prioritizing affective interpretability over error correction.

Similarly, Yi et al. proposed a “Complementary Fusion Framework” utilizing a Mixture of Experts (MoE) to dynamically weight modality contributions [4]. The success of such decomposed architectures highlights the limitations of naive early fusion (concatenation) in handling complex cross-modal dynamics, driving the field toward specialized sub-modules. However, unlike the opaque gating mechanisms found in MoE architectures, we investigate interpretable scalar injection to resolve specific modal ambiguities.

In contrast to these highly complex architectures, this thesis investigates whether a stronger *inductive bias*, specifically the explicit injection of interpretable psychometric dimensions (Arousal/Valence), can achieve competitive performance with greater transparency than latent cross-modal attention or mixture-of-experts approaches.

2.3 Dimensional Affect and the Modality Gap

While categorical emotion recognition remains the standard task in SER, the psychological literature consistently posits that these categories are emergent properties of fundamental continuous dimensions, primarily *Arousal* (intensity) and *Valence* (polarity/positivity) [3]. A persistent challenge in multimodal SER is the disparity in how these dimensions are encoded across modalities, often referred to as the *Modality Gap*.

2.3.1 The Acoustic-Semantic Dichotomy

Acoustic features, particularly prosody (pitch, energy, speech rate), are strongly correlated with Arousal. High-energy emotions like *Anger* and *Happiness* share similar acoustic profiles, often characterized by high fundamental frequency (F_0) and intensity. Consequently, unimodal acoustic models typically achieve high performance on Arousal regression but struggle to disambiguate Valence (e.g., distinguishing *Happy* from *Angry*) [13]. Wagner et al. demonstrated that the introduction of Transformer-based acoustic models (e.g., Wav2Vec 2.0) has begun to narrow this gap by implicitly capturing linguistic information, yet the “Valence Gap” remains a bottleneck for acoustic-only systems [13]. Conversely, semantic models are better at resolving Valence but often fail to capture Arousal, particularly when the emotional intensity is conveyed through voice quality rather than lexical choice (e.g., a sarcastic “That’s great”). While acoustic models traditionally struggle with Valence, Wagner et al. demonstrated that this “Valence Gap” can be effectively closed by fine-tuning the transformer layers of foundation models like Wav2Vec 2.0 [13]. They empirically showed that static embeddings alone fail to capture the subtle paralinguistic cues required for Valence detection, necessitating the adaptation of the encoder’s internal representations. In this study, we

leverage this adaptation by utilizing a Wav2Vec 2.0 checkpoint that has been explicitly fine-tuned on the MSP-Podcast emotion corpus [13]. However, to maintain a modular and lightweight architecture, we utilise these backbones as frozen feature extractors, training only the downstream fusion layers.

2.3.2 Granularity in Feature Extraction

To bridge this gap, recent work has moved beyond global utterance-level statistics toward fine-grained temporal modeling. Gallo-Aristizábal et al. demonstrated that extracting Wav2Vec 2.0 embeddings at the phoneme, syllable, and word levels yields superior granularity for distinguishing pathological speech patterns compared to global averaging [19]. This supports the hypothesis that word-level alignment, synchronizing acoustic frames with specific semantic tokens, is essential for capturing localized emotional cues that may be lost in global pooling.

2.3.3 Explicit vs. Implicit Integration Strategies

Recent state-of-the-art approaches have attempted to resolve the modality gap through complex architectural fusion. Gao et al. proposed a multi-level interaction mechanism that explicitly filters semantic features based on ASR error predictions [5]. Similarly, Yi et al. introduced a complementary fusion framework utilizing a Mixture of Experts MoE to dynamically weight modality contributions [4].

While these methods achieve high accuracy, they often rely on implicit latent alignment mechanisms that function as "black boxes," obscuring the decision boundary. In contrast to these high-complexity architectures, Parthasarathy and Busso demonstrated that jointly modeling Arousal and Valence as explicit multi-task targets improves the robustness of the representation [3]. Building on this finding, this thesis investigates whether the *explicit injection* of these psychometric primitives (e.g., concatenating a predicted Arousal Scalar to the fusion vector) provides a more interpretable and data-efficient alternative to implicit cross-modal attention for resolving the ambiguities between high-energy and low-Valence states.

2.4 Cross-Modal Attention Architectures

The dominant paradigm for resolving the modality gap in recent literature is the *Cross-Modal Attention* mechanism, first formalized in the MulT by Tsai et al. [18]. Unlike early fusion (concatenation) or late fusion (ensemble), MulT introduces a latent adaptation where streams from one modality (e.g., acoustics) are used to explicitly re-weight the features of another (e.g., text) via a query-key-value attention mechanism ($Q_{text}, K_{audio}, V_{audio}$). This allows the model to learn long-range dependencies across asynchronous streams without requiring rigid pre-alignment.

Building on this, recent works have introduced hierarchical variations to capture interactions at different temporal granularities. Dutta and Ganapathy proposed the Hierarchical Cross Attention Model (HCAM), which employs a curriculum learning strategy to fuse audio and text at the utterance level [20]. Similarly, Liu et al. introduced a HCMT for abnormal behavior recognition, fusing visual and audio snippets to capture global video features [21].

Positioning of Our Approach In this thesis, we adapt the cross-modal attention paradigm specifically for the task of word-level SER. We adopt the nomenclature *HCMT* to describe our architecture, which applies the cross-attention mechanism described by Tsai et al. [18]. We explicitly distinguish our implementation from the vision-centric HCMT proposed by Liu et al. [21]; while the nomenclature is shared, our architecture is distinct in its use of foundational speech (Wav2Vec 2.0) and language (BERT) models as the primary encoders, rather than the visual temporal segment transformers used in behavioral analysis. Our focus lies in evaluating whether this complex latent alignment offers superior interpretability compared to explicit psychometric injection (RQ3), rather than proposing a novel transformer block *ab initio*.

2.5 Explainability and Interpretability in Multimodal Fusion

In the context of Deep Learning (DL) for SER, it is necessary to distinguish between *post-hoc explainability*, which attempts to justify a decision after it has been made, and *intrinsic interpretability*, where the model architecture itself is constrained to operate on human-understandable concepts.

2.5.1 Post-hoc Explainability via Attention and Gating

The dominance of Transformer architectures [17] has normalized the use of attention weights as a proxy for feature importance. In multimodal settings, this involves visualizing cross-modal attention matrices to determine, for example, which text tokens trigger specific acoustic focus. Methods like the Interpretable Dynamic Fusion Graph [22] attempt to map these interactions explicitly by visualizing the dynamic efficacies of connections between modality vertices. However, raw attention weights are often criticized as insufficient explanations, as they represent statistical correlation within the high-dimensional latent space rather than causal logic.

A distinct evolution in this domain is the MAG, proposed by Rahman et al. [9]. Unlike pure attention mechanisms that re-weight features based on similarity, MAG computes a gating vector derived from non-verbal modalities to explicitly *shift* the internal semantic representations of pre-trained language models like BERT. This "gating" approach treats acoustics not merely as a context to be attended to, but as a modulation signal that displaces the semantic embedding space. In contrast, the HCMT architectures investigated in this thesis rely on standard scaled dot-product attention rather than gated shifting. We hypothesize that while gating may be efficient for large-scale pre-training, pure cross-attention offers a more granular view of word-level acoustic-semantic alignment, provided the attention mechanism does not collapse.

2.5.2 Taxonomy of Attention Mechanisms

To rigorously categorize the interaction mechanisms used in SER, Lieskovská et al. provide a taxonomy of attention based on the alignment scoring function [23]. They distinguish between *content-based* attention, where alignment scores are derived from the similarity of hidden states (e.g., matching a query vector to a key vector), and *location-based* attention, where alignment is determined solely by position within the sequence.

The cross-modal mechanism employed in our HCMT falls strictly under the **content-based** category. The semantic tokens serve as Queries (Q) and the acoustic frames as Keys (K); the attention weights are computed via the compatibility function $softmax(\frac{QK^T}{\sqrt{d_k}})$. This implies that the model associates modalities based on the feature content (e.g., the vector representation of the word "angry" matching the vector representation of high-energy prosody) rather than temporal proximity. This distinction is critical for our analysis of "recency bias" in later chapters: a failure of content-based attention to find semantic matches often results in a fallback to positional heuristics, effectively degenerating into a naive location-based mechanism.

2.5.3 Intrinsic Interpretability via Domain Concepts

While attention maps offer a window into the model's focus, they remain "black boxes" regarding the decision logic. An alternative approach is imposing domain-specific constraints on the architecture to achieve intrinsic interpretability. By explicitly modeling intermediate psychological dimensions, systems can offer explanations grounded in affective science rather than abstract tensor operations. For instance, conditioning a classifier on a predicted Arousal Scalar allows the model's decision boundary to be analyzed in terms of emotional intensity (e.g., distinguishing *Happiness* from *Neutrality* via energy levels) rather than opaque latent variance. This motivates the *Arousal-Augmented Fusion* strategy proposed in this work, which prioritizes explicit psychometric injection over complex latent entanglement.

3 Methods

3.1 Feature Extraction Pipeline

The foundation of our multimodal models is a robust feature extraction pipeline designed to generate parallel acoustic and semantic representations from the raw IEMOCAP data. To facilitate a comprehensive analysis as outlined in our research questions, we extract features at two distinct granularities: the entire utterance and the individual word level. This dual-level approach allows us to compare coarse, global representations against fine-grained, temporally-aligned sequences. This frozen feature extraction protocol ensures a lightweight training phase. Crucially, this introduces an asymmetry in our features: the acoustic backbone (Wav2Vec 2.0) utilises transfer learning from an emotion-specific domain (MSP-Podcast), while the semantic backbone (BERT) utilizes generic linguistic representations (Wikipedia/BooksCorpus) without task-specific fine-tuning.

3.1.1 Acoustic Representations

We extract two types of acoustic features to contrast traditional, theory-driven descriptors with modern, self-supervised representations.

eGeMAPS As a state-of-the-art, hand-crafted feature set, we use the eGeMAPS [6]. This set of 88 parameters is specifically designed for affective computing and offers a high degree of interpretability. Using the openSMILE toolkit [24], we extract:

- **Functionals**: A single 88-dimensional feature vector for each utterance, created by calculating summary statistics (e.g., mean, standard deviation) over the LLDs. This forms our utterance-level acoustic representation.
- **LLDs**: Frame-level features extracted at a 10ms hop size. These time-series features are later aligned to word boundaries to form word-level acoustic representations.

Wav2Vec 2.0 To leverage learned representations from raw audio, we employ the SSL paradigm. We utilized the Hugging Face *transformers* library [25] to instantiate the specific checkpoint *audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim* [13]. Unlike the standard Wav2Vec 2.0 Base model, which is trained solely on clean read speech (LibriSpeech), this "Robust" Large variant (317M parameters) was pre-trained on diverse domains, including Switchboard and Fisher, to handle variability in recording conditions.

Crucially, this specific checkpoint was subsequently fine-tuned on the MSP-Podcast corpus to predict dimensional emotion targets (Arousal, Valence, and Dominance). Consequently, the extracted embeddings are not merely acoustic descriptors but are explicitly optimized for paralinguistic tasks.

- For the **utterance-level** representation, we pass the entire audio signal through the model and apply temporal mean pooling over the final hidden states of the transformer encoder, yielding a single 1024-dimensional vector.
- For the **word-level** representation, we align the model's frame-level outputs with word timestamps derived from the forced alignment pipeline. The frames corresponding to each word are then mean-pooled to produce one 1024-dimensional vector per word.

Table 3.1 Comparison of Feature Units: Baseline (Utterance-Level) vs. Proposal (Word-Level). Note that for acoustic features, frame-level data is aggregated to match word boundaries.

(a) Stream A: Utterance Level (Global)			**(b) Stream B: Word Level (Sequential)**		
Modality	**Method**	**Shape**	**Modality**	**Method**	**Shape**
Acoustic	eGeMAPS Functionals	$1 \times D_{\text{func}}$	Acoustic	LLDs + Word Pooling	$W \times D_{\text{LLD}}$
Acoustic	Wav2Vec 2.0 (Mean)	$1 \times D_{\text{w2v}}$	Acoustic	Frames + Word Pooling	$W \times D_{\text{w2v}}$
Semantic	Word2Vec (Mean)	$1 \times D_{\text{w2v}}$	Semantic	Word2Vec	$W \times D_{\text{w2v}}$
Semantic	BERT [CLS]	$1 \times D_{\text{bert}}$	Semantic	Subwords $\rightarrow$ Word Mean	$W \times D_{\text{bert}}$

Notation: W denotes the number of words in the utterance. $D_{\text{func}} = 88$ (functionals), $D_{\text{LLD}} = 23$ (low-level descriptors), $D_{\text{w2v}} = 1024$ (Wav2Vec Large), $D_{\text{bert}} = 768$ (BERT Base).

3.1.2 Semantic Representations

Parallel to the acoustic features, we extract semantic embeddings from the provided transcripts.

BERT As our primary semantic feature extractor, we utilize the BERT [8]. We specifically selected the *google-bert/bert-base-uncased* checkpoint via the *transformers* library. This model consists of 12 transformer layers (110M parameters) and was pre-trained on the BooksCorpus and English Wikipedia. We selected the uncased version to align with the normalized, lower-case format of the IEMOCAP transcripts.

- For the **utterance-level** representation, we extract the embedding of the special *[CLS]* token from the final hidden layer, which provides an aggregate representation of the entire sequence ($d = 768$).
- For the **word-level** representation, we address the mismatch between BERT's subword tokenization (WordPiece) and linguistic words. We apply a "subword-to-word" pooling strategy: the embeddings of all sub-tokens constituting a single word are averaged to yield a single 768-dimensional vector per word.

3.1.3 Preprocessing and Temporal Alignment

A critical step in preparing the data for our models, especially for the word-level architectures, is the alignment and normalization of sequences.

Segmentation and Alignment The raw IEMOCAP session recordings are first segmented into discrete utterances using the provided annotations. To enable our word-level analysis, we then perform forced alignment on the audio and transcripts to obtain precise start and end timestamps for each word. These timestamps are the key to synchronizing the acoustic LLDs and Wav2Vec 2.0 frames with the corresponding BERT and Word2Vec embeddings, creating a parallel multimodal sequence for each utterance.

3.1.4 Formalization of Granularity and Pooling

To ensure reproducibility, we formally define the aggregation operations used to generate the feature vectors described in Table 3.1.

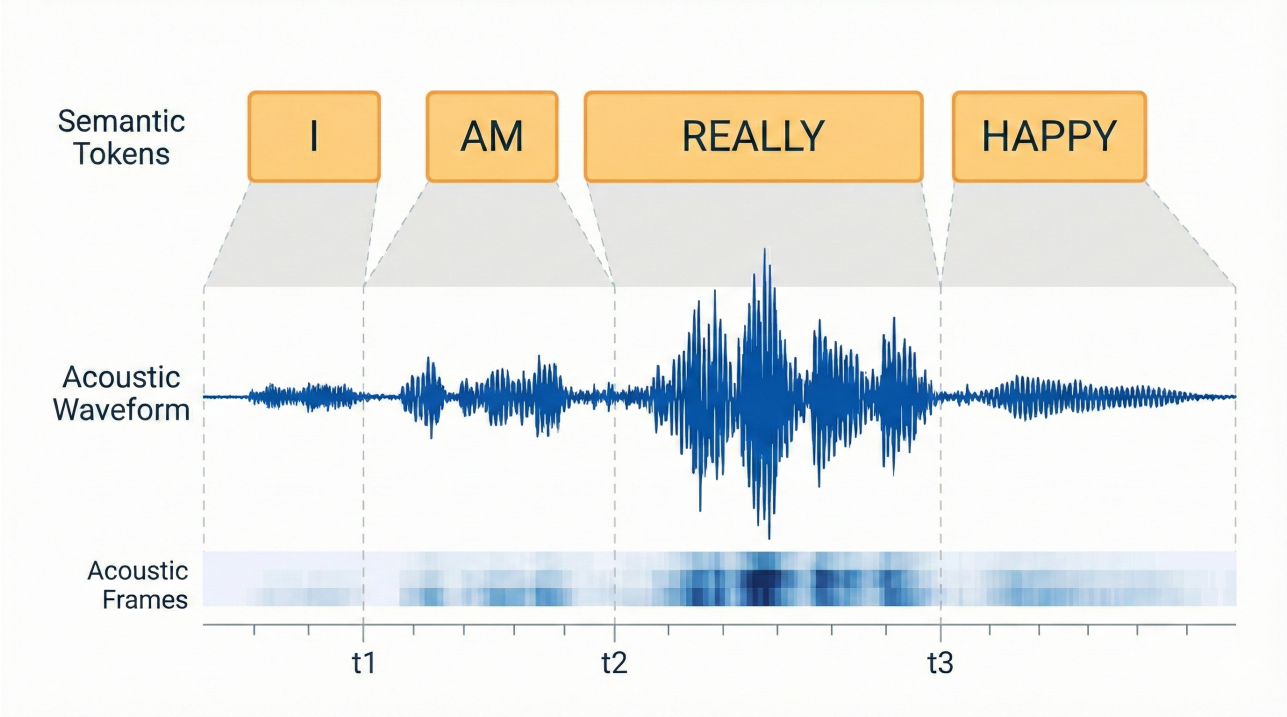


Figure 3.1 Schematic of Word-Level Forced Alignment. The diagram illustrates the synchronization of heterogeneous modalities. The **Semantic Stream** (top) consists of discrete lexical tokens, while the **Acoustic Stream** (bottom) represents the continuous audio waveform. Trapezoidal regions depict the alignment mapping, associating each token with a specific temporal span $[t_i, t_{i+1}]$ derived from forced alignment timestamps. This process enables the extraction of word-aligned acoustic features (e.g., mean-pooled Wav2Vec 2.0 frames) to match the granularity of the semantic embeddings. *Note: The utterance "I AM REALLY HAPPY" is a synthetic example constructed for visualization purposes and does not originate from the IEMOCAP dataset. Figure generated with AI (Google Gemini).*

Acoustic Word Alignment Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_T)$ be the sequence of acoustic frames (either eGeMAPS LLDs or Wav2Vec 2.0 frames), where T is the total number of frames in the utterance. Given the forced alignment timestamps for the i-th word w_i as start time $t_{start}^{(i)}$ and end time $t_{end}^{(i)}$, the word-level acoustic representation $\mathbf{x}_{ac}^{(i)}$ is computed via temporal mean pooling:

$$\mathbf{x}_{ac}^{(i)} = \frac{1}{K_i} \sum_{k \in \mathcal{T}_i} \mathbf{a}_k, \quad \text{where } \mathcal{T}_i = \{k \mid t_{start}^{(i)} \leq \text{time}(k) < t_{end}^{(i)}\} \tag{3.1}$$

where $K_i = |\mathcal{T}_i|$ is the number of frames corresponding to the word.

Semantic Subword Pooling BERT utilizes a WordPiece tokenizer that may split a single linguistic word into multiple sub-tokens. Let $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_M)$ be the sequence of subword embeddings generated by the BERT encoder. If the i-th word corresponds to the sub-token span $[j, k]$, the word-level semantic representation $\mathbf{x}_{sem}^{(i)}$ is derived by:

$$\mathbf{x}_{sem}^{(i)} = \frac{1}{k - j + 1} \sum_{n=j}^{k} \mathbf{s}_n \tag{3.2}$$

This ensures that both acoustic and semantic streams are synchronized to the same resolution W (words), enabling the multimodal sequence modeling in Stream B.

Sequence Normalization via Random Cropping Neural network models are typically trained on batches of fixed-size inputs. However, utterances in IEMOCAP have variable lengths, with a mean length of approximately 11.6 words. To address this, we employ a random cropping strategy during training. The selection of the crop size was determined through an empirical sensitivity analysis, as illustrated in Figure 3.2.

While performance gains are sharp as the crop size reaches the mean utterance length, we observed that the Unweighted Average Recall (UAR) reaches an optimal plateau and stabilizes within the range of 25 to 40 words. Beyond a crop size of 60, the variance increases significantly, likely due to the excessive amount of padding required for the majority of sequences. Consequently, we selected a fixed crop size of 30. This ensures that almost all utterances are captured in their entirety while maintaining high model stability and performance.

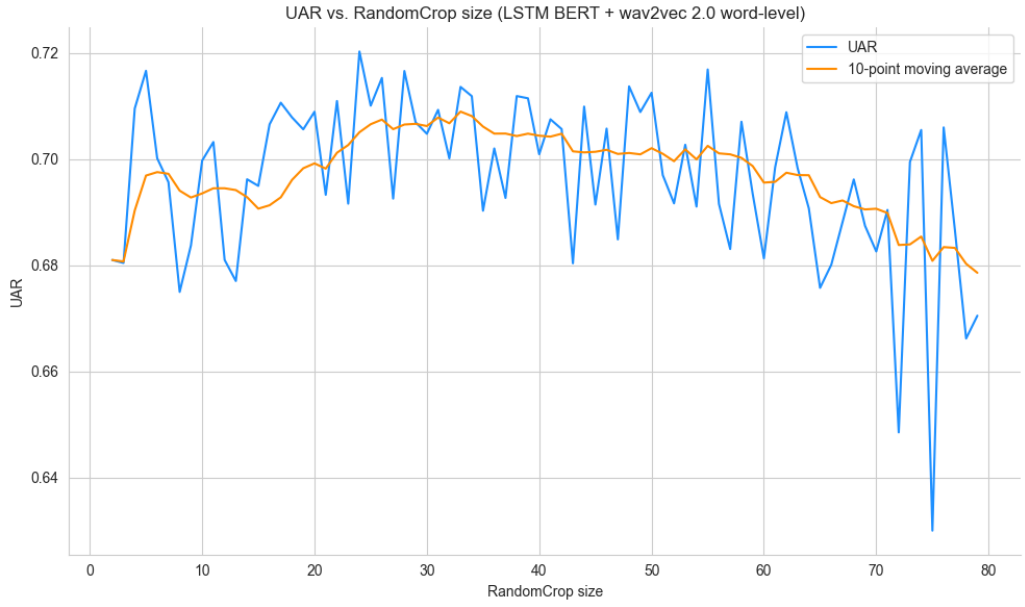


Figure 3.2 Effect of RandomCrop size on model performance (UAR). The 10-point moving average indicates optimal performance and stability in the 25 - 40 word range, followed by increased variance at higher crop sizes.

3.2 Experimental Design: The Granularity Contest

To address **RQ1**, we define two distinct modeling streams that isolate the effect of temporal resolution from model capacity:

- **Stream A: Global Utterance-Level Descriptors.** Input features are aggregated via mean-pooling (for BERT and Wav2Vec 2.0) or functional statistics (for eGeMAPS) into a single vector $v \in \mathbb{R}^d$ per utterance. These are processed by a Feed-Forward Neural Network (FFNN) classifier. This stream tests the hypothesis that emotion is a "global" property of the utterance.
- **Stream B: Fine-Grained Word-Level Sequences.** Input features preserve temporal resolution, aligned to word boundaries $w_{1...T}$. These sequences are processed by LSTM networks. This stream tests the hypothesis that emotion requires modeling the temporal evolution of acoustic and semantic cues.

3.2.1 Word-Level Early Fusion (Aligned LSTM)

To leverage the precise temporal alignment obtained via the forced alignment pipeline (Section 3.1.3), we implement a Word-Level Early Fusion model. Since both acoustic features $A = (a_1, \ldots, a_T)$ and semantic features $S = (s_1, \ldots, a_T)$ are aligned to the same word boundaries $t = 1 \ldots T$, we construct a fused sequence via simple concatenation:

$$x_t = \text{concat}(a_t, s_t) \in \mathbb{R}^{d_{acoustic}+d_{semantic}} \tag{3.3}$$

This sequence $X = (x_1, \ldots, x_T)$ is passed to a Bimodal LSTM (BiLSTM). This baseline tests the hypothesis that fine-grained, synchronized cross-modal interactions are critical for SER.

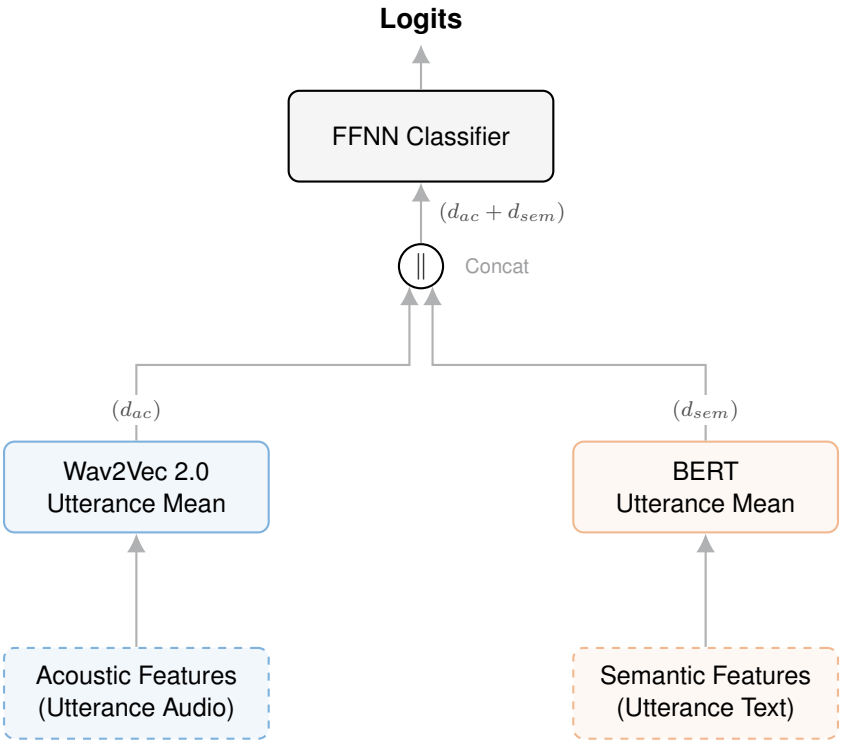


Figure 3.3 Architecture of the Baseline Early Fusion model. Utterance-level features (d_{ac} and d_{sem}) are processed through parallel paths and concatenated before the classification head.

3.3 Explicit vs. Implicit Integration Strategies

We investigate two distinct paradigms for integrating acoustic and semantic information.

Strategy I: Explicit Dimensional Augmentation (Late Fusion). This strategy operates on the hypothesis that categorical emotions in acted datasets are strongly correlated with psychometric dimensions, specifically Arousal (intensity) and Valence (positivity). We train a dedicated regressor R_θ to predict the Arousal score $\hat{a}$ from acoustic features alone.

Instead of standardizing the regression output, we apply a Sigmoid activation function to the final layer of the regressor. This design choice was motivated by the distribution of the target labels. By constraining the regressor output $\hat{a}$ to $(0, 1)$, we ensure the explicit signal lies within a bounded numerical range, theoretically stabilizing the gradients when concatenated with the high-dimensional latent embeddings:

$$\hat{a} = \sigma(\mathbf{W}^T \mathbf{h}_{acoustic} + b) = \frac{1}{1 + e^{-(\mathbf{W}^T \mathbf{h}_{acoustic} + b)}} \tag{3.4}$$

This scalar $\hat{a}$ is then concatenated with the frozen Wav2Vec 2.0 and BERT embeddings prior to the classification head, explicitly injecting domain knowledge into the decision boundary.

Strategy II: Latent Cross-Modal Attention (HCMT). To test if the model can learn these correlations implicitly, we implement a HCMT. Unlike simple concatenation, the HCMT utilizes a cross-attention mechanism where semantic tokens serve as queries (Q) and acoustic frames serve as keys (K) and values (V). As illustrated in Figure 3.7, this allows the semantic backbone to dynamically attend to acoustic events. This architecture theoretically offers higher expressivity but lacks the inductive bias provided by the explicit Arousal regressor.

3.4 Explicit Dimensional Augmentation Architecture

The Explicit Dimensional Augmentation strategy is designed to directly test the hypothesis that providing the model with an interpretable, domain-specific prior can resolve key ambiguities in emotion classification. The architecture, illustrated in Figure 3.4, consists of two distinct modules: a dedicated Arousal Regressor and an augmented Emotion Classifier.

Module A: The Arousal Regressor. First, an independent regression model is trained to map utterance-level acoustic features to a continuous Arousal score. This regressor is a simple FFNN that takes the 1024-dimensional mean-pooled output of a pre-trained Wav2Vec 2.0 model as its input. As described previously, a Sigmoid activation function is applied to the output layer to constrain the predicted Arousal score, $\hat{a}$, to the range $(0, 1)$. This module is trained separately on the continuous Arousal annotations provided in the IEMOCAP dataset.

Module B: The Augmented Emotion Classifier. The second module is the main emotion classifier. For each sample, we first extract the standard utterance-level embeddings from Wav2Vec 2.0 ($\mathbf{h}_{acoustic}$) and BERT ($\mathbf{h}_{semantic}$). We then use the pre-trained Arousal Regressor from Module A to predict the Arousal score $\hat{a}$ from the acoustic embedding. The final fusion vector, $\mathbf{v}_{fusion}$, is constructed by concatenating these three components:

$$\mathbf{v}_{fusion} = \mathbf{h}_{acoustic} \oplus \mathbf{h}_{semantic} \oplus \hat{a} \tag{3.5}$$

where $\oplus$ denotes concatenation. This augmented vector, which explicitly encodes emotional intensity alongside the rich latent representations, is then passed to a final FFNN classifier for emotion prediction. This design allows us to directly measure the impact of injecting an explicit psychometric primitive into the decision-making process.

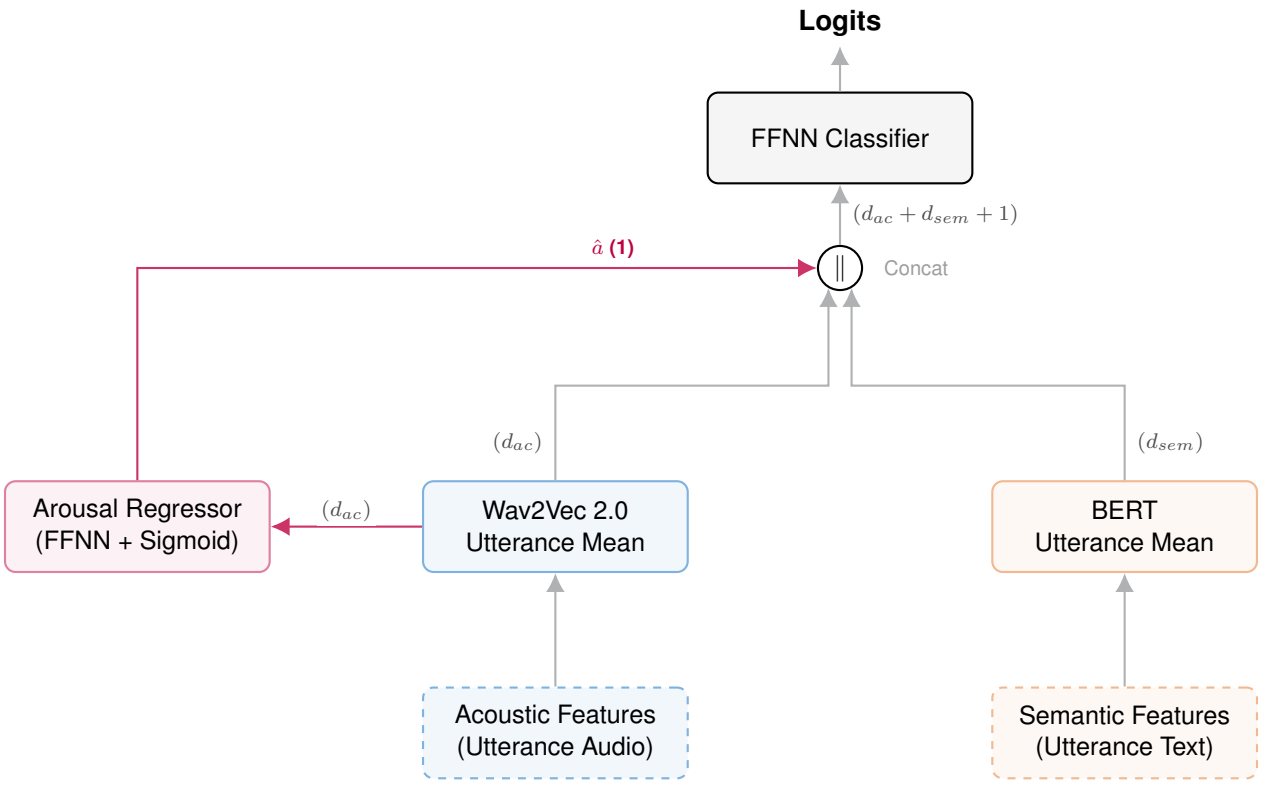


Figure 3.4 Architecture of the Arousal-Augmented Fusion model. A dedicated regressor predicts an arousal score ($\hat{a}$) from acoustic features. This scalar is then concatenated with the primary acoustic and semantic embeddings to explicitly inform the final classifier.

To empirically validate our hypothesis, we implemented a feature-fusion strategy. First, an independent regression model was trained on the IEMOCAP's continuous activation annotations to predict a single, utterance-level Arousal score ($\hat{a}$). This predicted scalar was then concatenated with the existing bimodal acoustic (Wav2Vec 2.0) and semantic (BERT) feature vectors for each sample. The resulting augmented feature vector provided the final emotion classifier with an explicit signal for emotional intensity, allowing us to directly test its influence on classification performance against the baseline model.

3.5 Latent Interaction Architectures (HCMT)

To investigate whether a model can implicitly learn the complex relationships between acoustic and semantic features, we implement a series of Transformer-based architectures inspired by the original design of Vaswani et al. [17]. These models leverage self-attention and cross-attention mechanisms to facilitate information fusion directly in the latent space. All variants share a common preliminary structure: separate modality-specific encoders process the acoustic and semantic sequences, which are then integrated using different fusion strategies. A special '[CLS]' token is prepended to the input sequences to serve as an aggregate representation for the final classification task, a common practice in Transformer-based models.

3.5.1 Dual-Encoder Transformer (Baseline)

The simplest variant, the Dual-Encoder Transformer, serves as our baseline for latent fusion. As depicted in Figure 3.5, this model processes the two modalities in complete isolation from each other. Each input sequence (acoustic and semantic) is passed through its own dedicated Transformer encoder. These encoders apply a series of self-attention layers, allowing each modality to build a context-aware representation based solely on its own temporal information.

Fusion occurs only after this modality-specific processing is complete. The final hidden state corresponding to the semantic '[CLS]' token is used as the summary representation for the text modality. For the acoustic modality, a global representation is obtained via masked average pooling over the final hidden states of its encoder. These two vectors are then concatenated and passed to a final FFNN for classification. This architecture represents a pure late-fusion strategy, as the modalities only interact at the final classification stage.

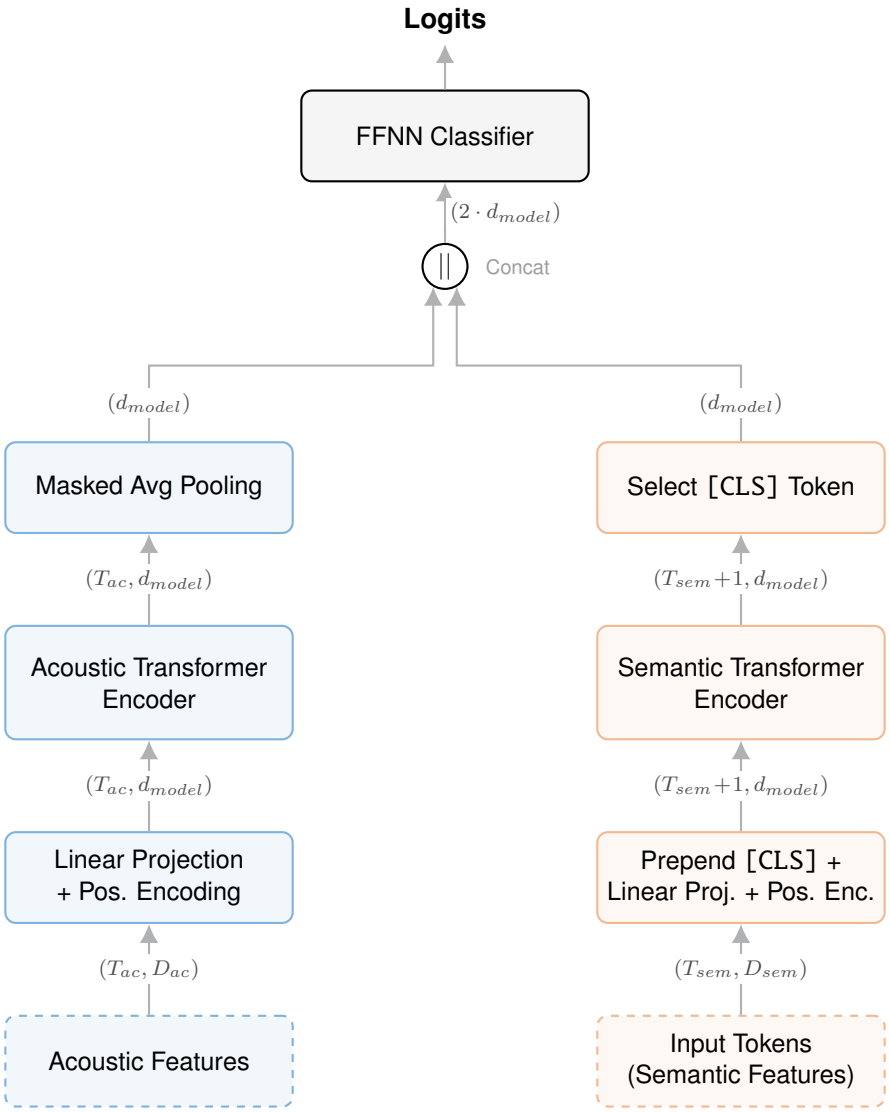


Figure 3.5 Architecture of the Dual-Encoder Transformer. The model processes acoustic (left) and semantic (right) streams independently before late fusion via concatenation.

3.5.2 HCMT-Unidirectional

The HCMT-Unidirectional model introduces a more sophisticated fusion mechanism by allowing for a one-way flow of information between modalities, as shown in Figure 3.6. After the initial modality-specific encoding via self-attention, a stack of cross-attention layers is introduced.

In this architecture, the semantic sequence acts as the query (Q) and the acoustic sequence provides the key (K) and value (V). This is implemented using a stack of *TransformerDecoderLayer* modules, where the semantic representations are passed as the target and the acoustic representations as the memory. This allows the model to refine the semantic representations by selectively attending to the most relevant acoustic features at each time step. The information flow is strictly unidirectional: from acoustics

to semantics. The final representation for classification is the hidden state of the [CLS] token from the output of this cross-modal fusion block.

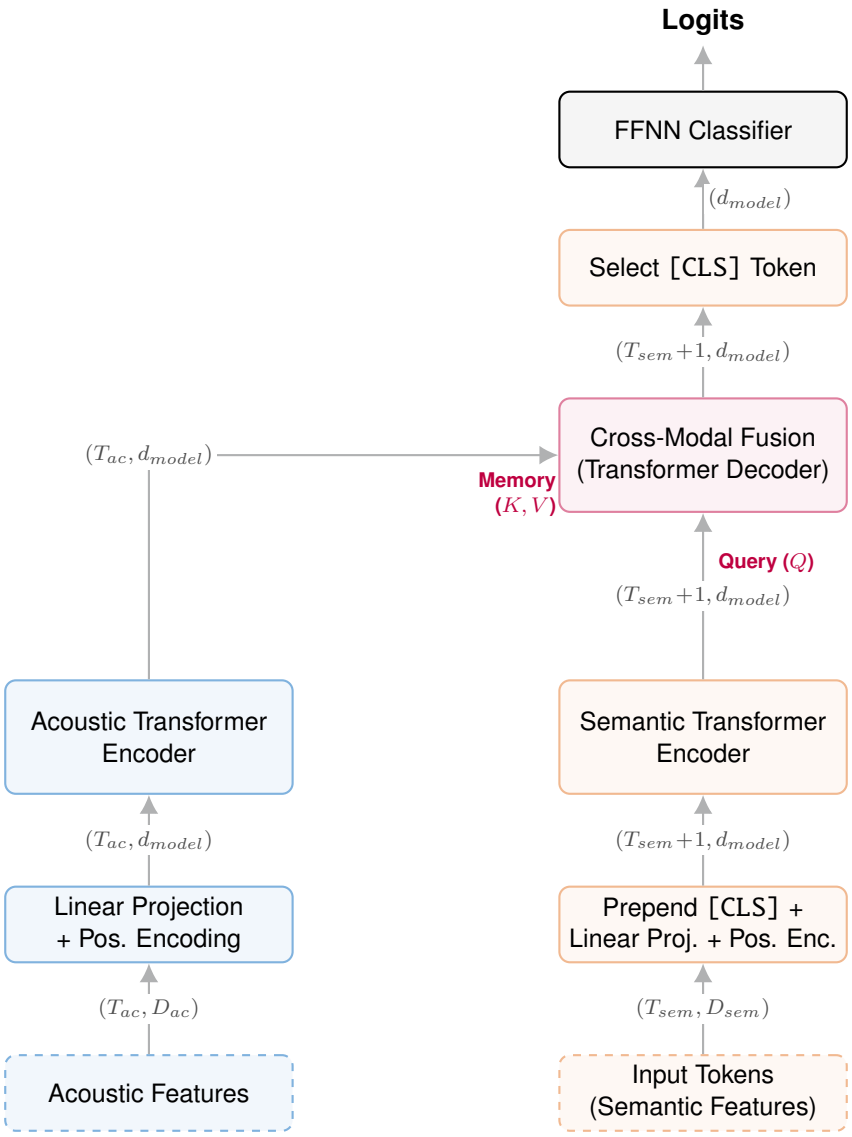


Figure 3.6 Architecture of the Hierarchical Cross-Modal Transformer (HCMTUnidirectional). Acoustic context (Memory) is dynamically injected into the semantic backbone (Query) via cross-attention.

3.5.3 HCMT-Bidirectional

The HCMT-Bidirectional model, illustrated in Figure 3.7, represents the most complex fusion strategy. It extends the unidirectional approach by creating a symmetric, two-way information exchange between the modalities. This is achieved by implementing two parallel cross-attention pathways:

1. **Acoustic-to-Semantic:** Identical to the unidirectional model, the semantic representations are refined by querying the acoustic sequence $(Q_{sem}, K_{ac}, V_{ac})$.
2. **Semantic-to-Acoustic:** Concurrently, the acoustic representations are refined by querying the semantic sequence $(Q_{ac}, K_{sem}, V_{sem})$.

This bidirectional flow allows for a deep, iterative fusion process where each modality can mutually inform the other. After the cross-modal fusion, the [CLS] token representations are extracted from the outputs of

both pathways. These two vectors are then concatenated to form the final feature vector, which is passed to the FFNN classifier. This architecture tests the hypothesis that a rich, bidirectional latent interaction is necessary to fully resolve cross-modal ambiguities.

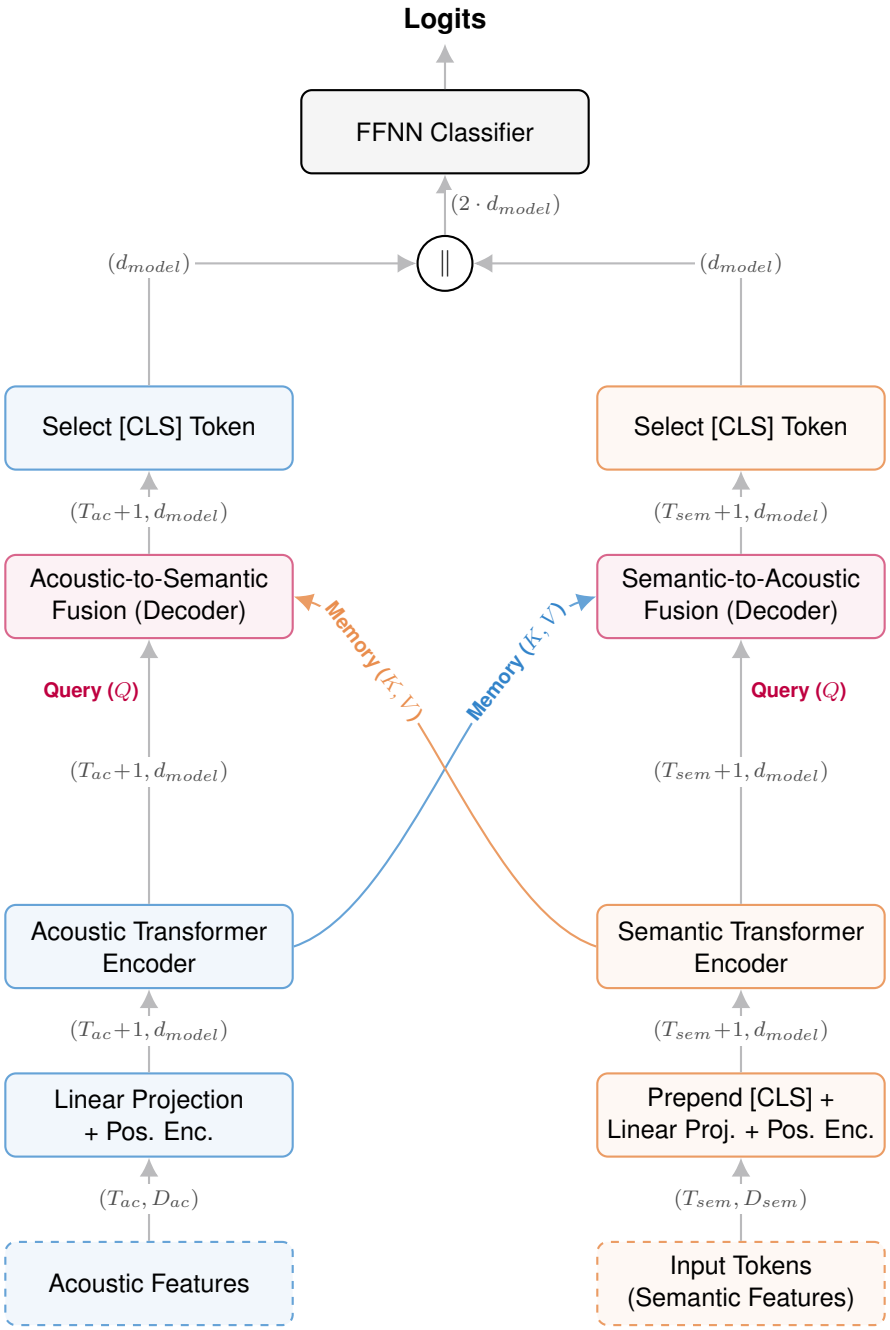


Figure 3.7 Architecture of the Hierarchical Cross-Modal Transformer (HCMTBidirectional). Both modalities serve as both Query and Memory in parallel decoder stacks, creating a symmetric fusion.

4 Experimental Setup

4.1 Dataset

To evaluate the proposed fusion architectures, we utilized the **IEMOCAP** database [10]. Collected at the University of Southern California, IEMOCAP is a standard benchmark in Speech Emotion Recognition SER designed to model expressive human communication in dyadic settings.

4.1.1 Corpus Description and Demographics

The dataset consists of approximately 12 hours of audiovisual data recorded from ten professional actors (5 male, 5 female). The interactions are organized into five discrete sessions. In each session, a distinct pair of actors (one male, one female) performs a series of dyadic interactions [10]. These interactions are divided into two elicitation paradigms:

1. **Scripted plays:** Actors memorize and perform selected emotional scripts.
2. **Improvised scenarios:** Actors improvise dialogue based on hypothetical scenarios designed to elicit specific affective states (e.g., *"telling a friend you are getting married"* to elicit happiness).

The reliance on professional actors in a studio environment ensures high audio quality and clear articulation. However, we must explicitly acknowledge a potential **"acting bias"** inherent to this collection method. As noted in recent literature [2], acted speech often relies on exaggerated prosodic cues, such as increased loudness and pitch variance to signal activation. This acting style may simplify the task of Arousal detection compared to naturalistic "in-the-wild" data.

4.1.2 Annotation and Class Filtering

The corpus contains detailed annotations at the utterance level. At least three human evaluators assessed each utterance using categorical labels chosen from a set of ten emotions: *Anger, Happiness, Sadness, Neutral, Frustration, Excited, Fear, Surprise, Disgust,* and *Other*. Additionally, dimensional annotations for *Valence*, *Activation* (Arousal), and *Dominance* were provided on a 5-point Likert scale by at least two evaluators [10].

Consistent with standard SER literature [5], we reduced the ten-class taxonomy to a four-class problem to address class sparsity and ensure varying degrees of Arousal and Valence are represented. The final target classes are **Anger, Happiness, Neutral, and Sadness**.

Handling the Happiness–Excitement Overlap A known issue in IEMOCAP is the semantic and acoustic overlap between *Happiness* and *Excitement*. Samples labeled as *Excited* are frequently merged into the *Happiness* class to balance the distribution against the dominant *Neutral* class [5]. Our analysis of the metadata confirms that *Happiness* alone accounts for a small fraction of the data. Therefore, we merged *Excited* into the *Happiness* category.

Data Selection Statistics We filtered the dataset to include only utterances with majority agreement for *Anger, Sadness, Neutral,* and the merged *Happiness/Excited* class. Based on our preprocessing of the provided metadata:

- The full dataset contains **10,039** samples with a total duration of approximately **12.4 hours** (44,775 seconds).
- After filtering for the four target classes, the experimental subset consists of **5,531** samples with a total duration of approximately **7.0 hours** (25,161 seconds).

The resulting class distribution typically exhibits a slight imbalance. As calculated in our analysis, the final distribution is: *Neutral* (30.9%), *Happiness+Excited* (29.6%), *Anger* (19.9%), and *Sadness* (19.6%).

4.2 Experimental Protocol

4.2.1 Partitioning and Speaker Independence

To ensure speaker independence, a critical requirement for avoiding identity confounds in SER, we adopted a session-based split strategy. Since each session contains unique speakers who do not appear in other sessions, splitting by session guarantees that the model is evaluated on unseen speakers.

We utilized the following fixed split protocol:

- **Training:** Sessions 1, 2, and 3.
- **Validation:** Session 4 (used for hyperparameter tuning and checkpoint selection).
- **Testing:** Session 5 (hold-out set for final evaluation).

While standard benchmarks on IEMOCAP typically employ 5-fold Leave-One-Session-Out (LOSO) cross-validation to mitigate speaker variance [5, 13], the high dimensionality of the extracted features and the computational overhead of processing 5,531 sequences through dual-stream transformers necessitated a fixed-split approach for this study. Furthermore, the fixed Session 5 split allows for direct comparability with recent literature utilising this specific protocol [13].

Prevention of Data Leakage in Dimensional Augmentation A critical methodological constraint was ensuring that the auxiliary regressors (used for the Arousal-Augmented Fusion) did not leak test-set information into the classifier. We applied the **identical split protocol** to the regressor training. The Arousal and Valence regressors were trained strictly on Sessions 1, 2, and 3. Consequently, the scalar Arousal values concatenated to the test set features (Session 5) were generated by a regressor that had never encountered the acoustic properties of the speakers in Session 5.

4.2.2 Training Framework and Hyperparameters

All models were implemented using the *autrainer* framework [15], which provides a standardized interface for training and evaluation on IEMOCAP.

Optimization We utilized the Adam optimizer [26] for all experiments. The learning rate was a critical hyperparameter, and a sweep was performed to find the optimal value for each class of model. Our empirical results, detailed in the master results table, revealed the following configuration yielded the best performance:

- **Multimodal Transformer Architectures (DualEncoderTransformer, HCMT):** The best results were consistently achieved with a learning rate of $\mathbf{10^{-4}}$. This rate proved sufficient to stabilize the complex cross-modal attention mechanisms without requiring a learning rate scheduler or warmup period.
- **Unimodal Baselines & FFNN Models:** These simpler models generally performed best with higher learning rates, typically in the range of 10^{-3} to 10^{-4}.

All training was conducted with a batch size of 32. We employed early stopping with a patience of 10 epochs, monitoring the UAR on the validation set (Session 4) to select the best-performing checkpoint for final evaluation.

4.2.3 Evaluation Metrics

Given the significant class imbalance in the IEMOCAP dataset, where *Neutral* and *Happiness* samples outnumber *Anger* and *Sadness*, standard accuracy is often a misleading indicator of performance. A model could achieve high accuracy by overfitting to the majority classes while failing to recognize minority emotions. To ensure scientific rigor and comparability with the state-of-the-art SER literature [5, 23, 13], we report three complementary metrics: UAR, Weighted Accuracy (WA), and Macro-F1 Score.

Unweighted Average Recall Our primary optimization and evaluation metric is UAR. It is defined as the arithmetic mean of the class-specific recall scores. This metric treats all emotion classes equally, regardless of the number of samples in the test set, preventing the majority class from dominating the evaluation.

$$\text{UAR} = \frac{1}{K} \sum_{k=1}^{K} \text{Recall}_k = \frac{1}{K} \sum_{k=1}^{K} \frac{TP_k}{TP_k + FN_k} \tag{4.1}$$

where $K = 4$ is the number of emotion classes, TP_k are the true positives, and FN_k are the false negatives for class k.

Weighted Average For completeness and comparison with general machine learning benchmarks, we report standard Accuracy, often referred to in SER literature as WA [23]. This measures the ratio of total correct predictions to the total number of samples N. Unlike UAR, this metric is biased toward the majority classes.

$$\text{WA} = \frac{\sum_{k=1}^{K} TP_k}{N} \tag{4.2}$$

Macro-F1 Score To provide a balanced view of precision and recall, we report the Macro-F1 score. This is calculated as the unweighted mean of the F1-scores for each class. It is particularly useful for analyzing the trade-offs between false positives and false negatives, as observed in our error analysis (e.g., the trade-off between *Neutral* precision and *Sadness* recall).

$$\text{F1}_{\text{macro}} = \frac{1}{K} \sum_{k=1}^{K} 2 \cdot \frac{\text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k} \tag{4.3}$$

5 Results and Analysis

5.1 Unimodal Baselines

To establish a baseline for multimodal integration, we first evaluate the predictive power of each modality in isolation. This analysis serves two purposes: first, it justifies the selection of high-capacity foundation models (Wav2Vec 2.0 and BERT) over traditional feature extraction methods (eGeMAPS and Word2Vec); second, it identifies the specific error modes inherent to acoustic and semantic channels, providing the context necessary to understand the fusion results discussed later in Section 5.2.

5.1.1 Acoustic Modality: Hand-Crafted vs. Self-Supervised

We compared the performance of hand-crafted features against self-supervised representations. As shown in Table 5.2, the SSL approach yields a substantial performance increase.

- **eGeMAPS (Baseline):** The classifier trained on the 88 functional features of the eGeMAPS set achieved a UAR of **0.548**. While computationally efficient, this shallow representation fails to capture the complex temporal dynamics of emotional expression in the IEMOCAP dataset.
- **Wav2Vec 2.0 (Foundation Model):** Fine-tuning the Wav2Vec 2.0 encoder followed by mean pooling achieved a UAR of **0.698**, an absolute improvement of +15.0% over the hand-crafted baseline.

This result confirms that deep acoustic embeddings extracted from raw waveforms capture paralinguistic information significantly better than expert-defined features. Notably, the acoustic model demonstrates exceptional sensitivity to high-Arousal emotions. An analysis of the confusion patterns reveals that the Wav2Vec 2.0 model achieves a recall of **0.91** for *Anger* (296/327 correctly classified), serving as the strongest single-modality predictor for this class. However, it exhibits a characteristic "acoustic confusion" between *Happiness* and *Anger*, misclassifying 24% of *Happy* utterances as *Angry*, likely due to their shared high acoustic energy (Arousal).

5.1.2 Semantic Modality: Static vs. Contextual Embeddings

In the semantic domain, we observed a similar performance gap between static word embeddings and transformer-based models.

- **Word2Vec:** The baseline model using global mean-pooling of Word2Vec embeddings achieved a UAR of **0.573**.
- **BERT:** The BERT model (using the *[CLS]* token) achieved a UAR of **0.641**, outperforming Word2Vec by +6.8%.

While the semantic model underperforms the acoustic model on aggregate metrics (UAR 0.641 vs 0.698), it provides complementary discriminative power. Specifically, BERT aids in distinguishing emotions with similar Arousal but different Valence. However, a detailed error analysis reveals that the semantic model struggles with the *Neutral* class, frequently misclassifying neutral statements as *Happiness* (33% error rate). This suggests that many "neutral" utterances in IEMOCAP contain positive lexical content (e.g., polite agreement) that lacks the prosodic flatness typically associated with neutrality, a nuance lost when ignoring the acoustic channel.

Table 5.1 Comprehensive performance comparison on the IEMOCAP test set (Session 5). Models are ordered by relevance, starting with the proposed Explicit Injection framework. The **Relative** column (Δ) indicates the relative improvement in UAR compared to the standard Multimodal Early Fusion baseline. The **Arousal-Augmented Fusion** achieves the highest UAR, outperforming both the baseline and the more computationally complex HCMT architectures.

Model Architecture	UAR	WA (Acc)	Macro-F1	Δ vs Baseline
Proposed Explicit Injection (Late Fusion)				
Arousal-Augmented Fusion	**0.743**	**0.729**	**0.730**	**+1.6%**
Valence-Augmented Fusion	0.742	0.728	0.730	+1.5%
Latent Alignment Architectures				
HCMT (Unidirectional)	0.737	0.712	0.716	+1.0%
Dual-Encoder Transformer	0.737	0.720	0.727	+1.0%
HCMT (Bidirectional)	0.721	0.708	0.709	-0.6%
Multimodal Baseline				
Early Fusion (Concat)	0.727	0.712	0.716	*Reference*
Unimodal Baselines				
Acoustic Only (Wav2Vec 2.0)	0.698	0.666	0.674	-2.9%
Semantic Only (BERT)	0.641	0.637	0.626	-8.6%

It is important to note that the BERT model achieves this performance (0.641 UAR) using only frozen, generic embeddings. Unlike the acoustic model, which benefits from prior fine-tuning on emotional data, the semantic model relies entirely on general-purpose linguistic knowledge, yet still provides competitive discriminative power.

Table 5.2 Comparison of Unimodal Baselines on IEMOCAP (Session 5). Foundation models (Wav2Vec 2.0 and BERT) significantly outperform traditional features (eGeMAPS and Word2Vec). The Acoustic modality is the dominant single channel for emotion recognition in this dataset.

Modality	Feature Extractor	UAR	Accuracy (WA)
Acoustic	eGeMAPS (Functionals)	0.548	0.504
	Wav2Vec 2.0 (Mean)	**0.698**	**0.666**
Semantic	Word2Vec (GoogleNews)	0.573	0.551
	BERT (Base Uncased)	0.641	0.637

5.2 Multimodal Early Fusion (Baseline)

To establish a robust baseline for multimodal integration, we evaluated an Early Fusion architecture that concatenates the global utterance-level representations from the acoustic encoder (Wav2Vec 2.0) and the semantic encoder (BERT) prior to classification. This approach tests the hypothesis that a naive linear combination of high-level features is sufficient to capture cross-modal dependencies.

5.2.1 Performance Overview

The Early Fusion model achieved a UAR of **0.727** and a Weighted Accuracy (WA) of **0.712** on the test set (Session 5) (see Table 5.1). This represents a substantial empirical improvement over both unimodal baselines. Specifically, the fusion model surpasses the acoustic-only baseline (0.698 UAR) by absolute **+2.9%** and the text-only baseline (0.641 UAR) by **+8.6%**.

This synergistic gain confirms that the two modalities provide complementary affective signals. As visualized in Figure 5.1, the fusion mechanism successfully mitigates the weaknesses of the individual encoders, particularly in cases where semantic ambiguity is resolved by prosodic cues.

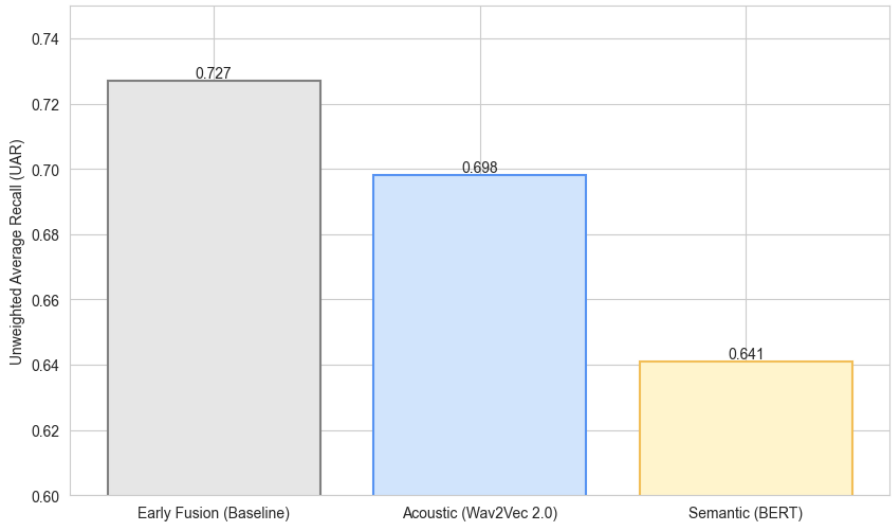


Figure 5.1 The Multimodal Synergy. Comparison of Unimodal Baselines against the Early Fusion Baseline. The fusion of Acoustic (Wav2Vec 2.0) and Semantic (BERT) features yields a +2.9% improvement in UAR over the strongest unimodal predictor.

5.2.2 Error Analysis: The Limits of Concatenation

Despite the performance gains, an analysis of the confusion matrix (Figure 5.2) reveals a critical limitation in the concatenation strategy. While the model achieves high recall for *Anger* and *Sadness*, it exhibits significant confusion between *Happiness* and *Neutral* states.

- **The Happy $\rightarrow$ Neutral Collapse:** As shown in Figure 5.2, 87 utterances labeled as *Happy* were misclassified as *Neutral*. This suggests that while semantic content helps distinguish Valence (e.g., positive words), the naive concatenation fails to effectively utilize acoustic intensity (Arousal) to distinguish the active state of *Happiness* from the passive state of *Neutrality*.
- **Sadness Precision:** Conversely, the fusion model performs robustly on *Sadness* (166 correct), suggesting that the acoustic markers for sadness (lower energy, slower tempo) are preserved effectively during concatenation.

This failure to distinguish *Happy* from *Neutral* highlights the "Modality Gap" described in Section 1.2. The concatenation operation treats acoustic and semantic features as static vectors, failing to explicitly model the *intensity* dimension required to separate these classes. This limitation motivates the *Arousal-Augmented Fusion* proposed in the subsequent section.

5.3 Granularity Analysis (Addressing RQ1)

To address **RQ1** ("Do word-level descriptors outperform utterance-level descriptors?"), we evaluate two distinct architectural paradigms across both unimodal and multimodal settings.

1. **Utterance-Level (Global):** Features are aggregated via mean-pooling (for Wav2Vec 2.0 and BERT) or functionals (for eGeMAPS) and processed by a Feed-Forward Neural Network (FFNN). This represents a "global context" approach.

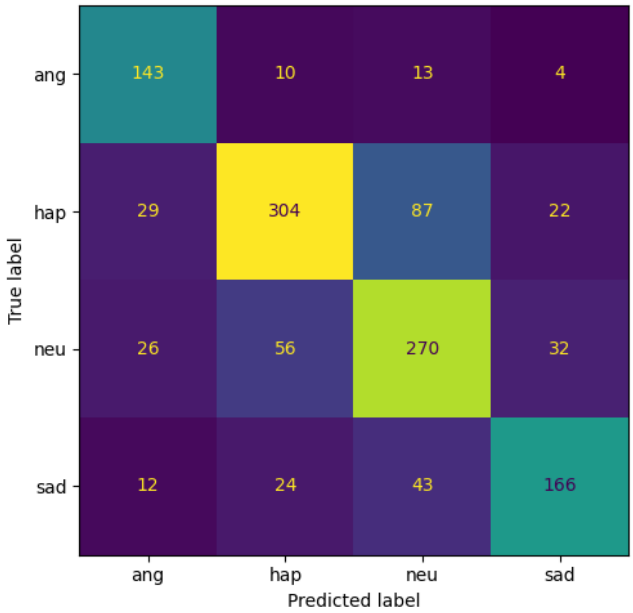


Figure 5.2 Confusion Matrix of the Early Fusion Baseline. The model struggles to distinguish high-energy positive emotions from neutral states, evidenced by 87 *Happy* samples misclassified as *Neutral* (Recency Bias/Intensity mismatch).

2. **Word-Level (Sequential):** Features are aligned to word boundaries and processed as a sequence $t = 1 \dots T$ using Long Short-Term Memory (LSTM) networks. This represents a "temporal dynamics" approach.

Table 5.3 presents the comparative performance.

Table 5.3 Impact of Granularity on Model Performance (UAR). The table compares **Stream A** (Global Utterance-Level features processed by FFNN) against **Stream B** (Sequential Word-Level features processed by LSTM). While acoustic modeling benefits slightly from sequential granularity, semantic and multimodal performance degrades significantly, suggesting that emotion in this dataset is a global property rather than a sequence of word-level events.

Modality	**Utterance-Level (FFNN)**	**Word-Level (LSTM)**	**Δ (Seq. vs Global)**
Unimodal Performance			
Acoustic (Wav2Vec 2.0)	**0.698**	0.678	-2.0%
Semantic (BERT)	0.641	**0.642**	+0.1%
Multimodal Fusion			
BERT + Wav2Vec 2.0	**0.727**	0.705	**-2.2%**

The Multimodal Alignment Cost: Crucially, when fusing modalities, the word-level approach significantly underperforms. Despite utilizing precise forced alignment to synchronize acoustic and semantic frames at the word level, the *lstm-bert+wav2vec* model achieves only 0.669 UAR, whereas the simple global concatenation (*ffnn-bert+wav2vec*) achieves 0.727.

Conclusion on RQ1: While word-level descriptors can compete in unimodal tasks, the computational effort and complexity required to align and model them sequentially is **not justified** for multimodal fusion in this dataset. The significant drop in bimodal performance suggests that the emotional content in text and speech is not strictly synchronized at the word boundary (e.g., a sarcastic tone may span the whole sen-

tence while the cue word is local). Consequently, global utterance-level fusion is not only more efficient but also more robust, serving as the superior baseline for the subsequent Arousal Augmentation experiments.

5.4 Effectiveness of Explicit Dimensional Augmentation

The Explicit Dimensional Augmentation strategy yielded the highest overall performance. Integrating **Valence** (UAR 0.742) or **Arousal** (UAR 0.743) improved upon the Early Fusion baseline (0.727) by approximately 1.6%. This confirms our hypothesis that injecting domain-specific psychometric priors is a highly efficient fusion strategy, surpassing even complex latent attention mechanisms.

5.4.1 Arousal: Resolving the Happiness-Neutral Ambiguity

The primary contribution of the Arousal signal was the disambiguation of high-energy positive emotions from neutral speech. Quantitative analysis of the test set reveals a distinct separation margin established by the regressor. While the model predicted a mean Arousal of $\mu = 0.06$ for *Neutral* samples, it predicted $\mu = 0.28$ for *Happy* samples. This constitutes an approximately five-fold increase in signal intensity.

As illustrated in the differential confusion matrices (Figure 5.5) and quantified in Table 5.4, this contrast allowed the fusion model to rectify the semantic ambiguity. While the semantic stream (BERT) provides a high probability mass for neutral/positive Valence, the injection of the predicted Arousal Scalar $\hat{a}$ acts as an intensity gate. The model learned to inhibit the *Neutral* class activation when $\hat{a}$ exceeded a learned threshold, thereby reallocating the probability mass to the high-Arousal *Happiness* class.

5.4.2 Valence: The Paradox of Uncalibrated Utility

A critical finding of this study is the success of Valence-Augmented Fusion (UAR 0.742) despite the statistical failure of the underlying regression module. As detailed in Section 5.4.3, the Valence regressor suffered from *negative collapse*, compressing predictions for *Happiness* ($\mu_{true} = 0.66$) into the lower tail of the Sigmoid function ($\mu_{pred} = 0.19$).

However, the classification results indicate that this "failed" regressor still improved the UAR by 1.4% over the baseline. This phenomenon is explained by distinguishing between *calibration* and *discrimination*:

- **Calibration Failure:** The regressor failed to map the absolute magnitude of Valence, effectively collapsing the output space to $[0.0, 0.2]$.
- **Discriminative Success:** Despite the compression, the Pearson correlation of $r = 0.40$ confirms that the *ordinality* was preserved. The model consistently ranked $Valence(Happy) > Valence(Neutral)$ (0.19 vs 0.03).

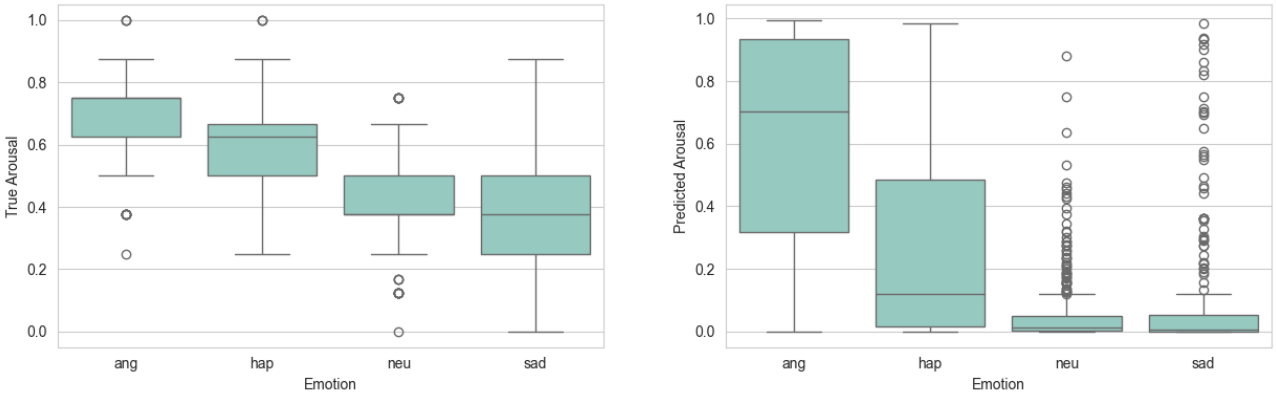
Since the downstream fusion classifier is scale-invariant regarding input features, it successfully learned to rescale this minute signal difference. This suggests that for SER classification, the *disentanglement* of polarity into a single monotonic scalar is more valuable than perfect absolute calibration.

5.4.3 Analysis of Dimensional Predictions

To validate whether the performance gain was indeed driven by accurate dimensional modeling, we analyzed the raw predictions of the regression modules on the test set. A visual inspection of the predictions reveals a systemic issue: both regressors exhibit significant *negative bias* and *scale compression*, likely induced by the saturating nature of the Sigmoid activation function and class imbalance. However, the functional utility of these compressed signals differs markedly between dimensions.

Arousal: Compression with Separability

The Wav2Vec 2.0-based Arousal regressor achieved a Pearson correlation coefficient of $r = 0.67$. As illustrated in Figure 5.3, comparing the ground truth against the predictions reveals the nature of the model's bias.



(a) Ground Truth. Note the full use of the scale $(0.0–1.0)$ and high variance.

(b) Predicted. The scale is compressed (< 0.5), but the relative ordering is preserved.

Figure 5.3 Arousal Distribution Comparison. The regressor suffers from negative bias (scale compression) induced by the Sigmoid activation. However, the *relative margin* between 'Happy' (hap) and 'Neutral' (neu) remains distinct in both plots, allowing the fusion model to use this signal as a discriminator.

Despite the compression shown in Figure 5.3b, the model establishes a clear *relative* margin.

- **Happiness:** Median prediction ≈ 0.15 (Ground Truth ≈ 0.6).
- **Neutral:** Median prediction ≈ 0.02 (Ground Truth ≈ 0.4).

While these values are numerically low, the separation allows the downstream classifier to treat the Arousal Scalar as a reliable "energy gate." The predicted signal for Happiness is consistently distinct from the noise floor occupied by Neutral and Sadness samples.

Valence: The Limit of Detectability

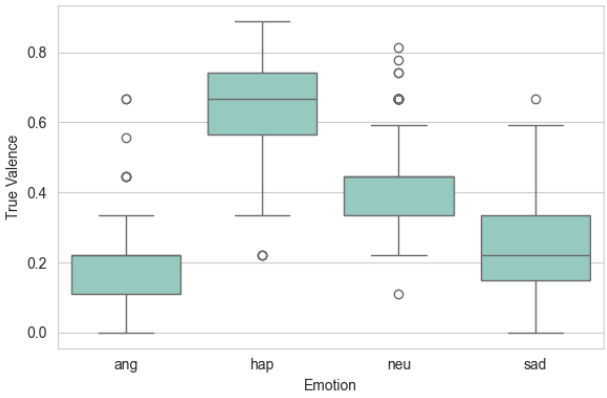
In contrast, the Valence regressor pushes the limits of detectability. Figure 5.4 compares the ground truth distributions against the model predictions, visualizing the phenomenon of *negative collapse*.

- **Happiness:** Mean predicted Valence $\mu = 0.19$ (Ground Truth $\mu = 0.66$).
- **Neutral:** Mean predicted Valence $\mu = 0.03$ (Ground Truth $\mu = 0.43$).

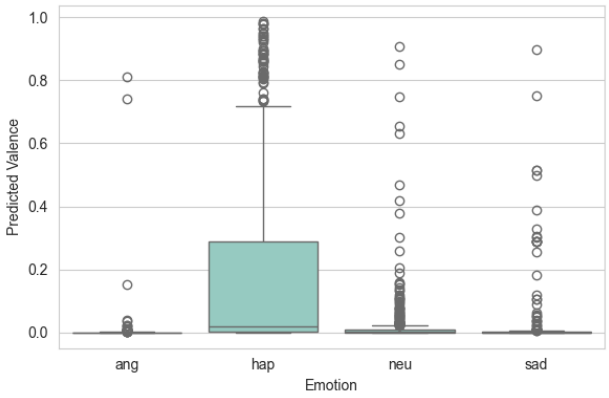
While the Pearson correlation ($r = 0.40$) confirms that an ordinal ranking exists, the signal operates on a micro-scale. The fusion classifier's success (UAR 0.742) implies it learned to apply extreme synaptic weights to this minute difference (0.19 vs 0.03). This confirms that while both dimensions suffered from calibration failure, Arousal retained a robust dynamic range, whereas Valence barely survived the collapse into the lower bound.

5.4.4 Mechanisms of Improvement: Arousal Gating vs. Valence Exclusion

A critical finding of this study is the performance parity between the Arousal-Augmented model (UAR 0.743) and the Valence-Augmented model (UAR 0.742), despite the statistical failure of the underlying Valence regressor (CCC $\approx$ 0.05). Analyzing the shift in error distributions reveals that these two dimensions optimize the decision boundary through fundamentally different mechanisms.



(a) Ground Truth. 'Happiness' (hap) is clearly separated ($\mu \approx 0.65$) from negative classes.



(b) Predicted. All classes collapse to the lower bound. 'Happiness' retains only minute variance.

Figure 5.4 Valence Distribution Comparison. The contrast highlights the calibration failure. While the ground truth shows a distinct polarity gap, the regressor compresses 'Happiness' to $\mu = 0.19$, barely distinguishing it from the noise floor of 'Neutral' ($\mu = 0.03$).

The Valence Paradox: Uncalibrated Utility

As detailed in Section 5.4.3, the Valence regressor suffered from *negative collapse*, compressing predictions for *Happiness* into the lower tail of the sigmoid function ($\mu_{pred} \approx 0.19$ vs. $\mu_{true} \approx 0.66$). However, the classification success implies a distinction between *calibration* and *discrimination*:

- **Calibration Failure:** The regressor failed to map the absolute magnitude of Valence, rendering the Mean Squared Error (MSE $\approx$ 0.054) high relative to the compressed output range.
- **Discriminative Success:** Despite the compression, the Pearson correlation ($r = 0.40$) indicates that ordinality was preserved. The fusion classifier, being scale-invariant regarding input features, successfully learned to weight this minute signal difference (0.19 vs. 0.03) to distinguish polarity.

Differential Error Correction

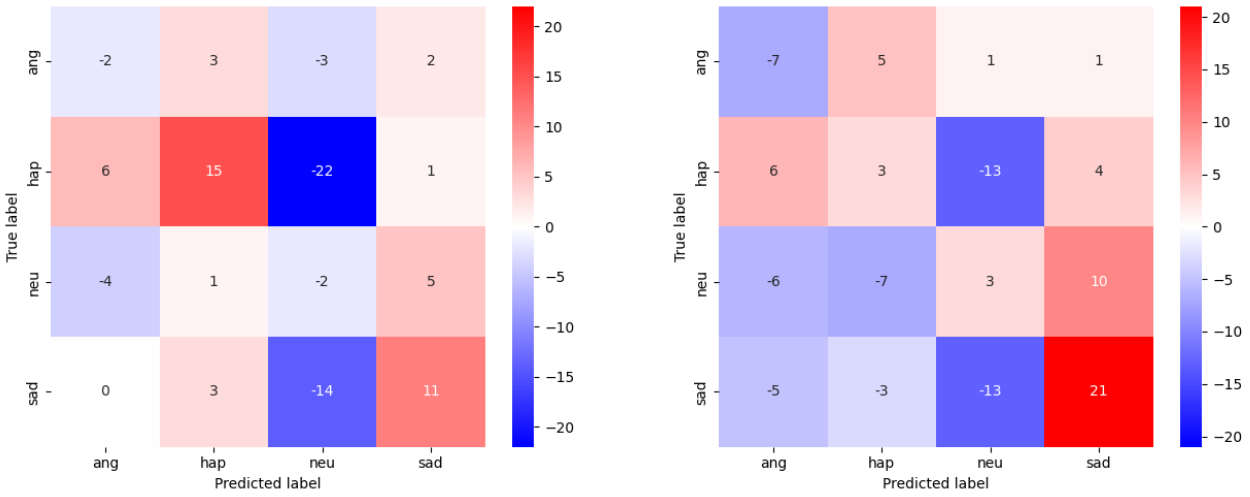
To understand the functional impact of these signals, we analyzed the changes in the confusion matrices relative to the Early Fusion baseline. Table 5.4 summarizes the class-specific improvements.

Table 5.4 Impact of Dimensional Augmentation on Class-Specific Recall. While Arousal improves the detection of active emotions (Happy), Valence improves the detection of passive negative emotions (Sadness) by filtering out positivity.

Metric	Baseline	+ Arousal	+ Valence
Happy Correct	304	**319** (+15)	307 (+3)
Happy $\rightarrow$ Neutral Error	87	**65** (-22)	74 (-13)
Sadness Correct	166	177 (+11)	**187** (+21)
Sadness $\rightarrow$ Neutral Error	43	29 (-14)	**30** (-13)

The **Arousal Regressor** functioned as an intensity *gate*. By establishing a clear signal margin between high-energy 'Happiness' and low-energy 'Neutrality', it significantly reduced the *Happy* $\rightarrow$ *Neutral* misclassification rate (-25.3%).

In contrast, the **Valence Regressor** functioned as an *exclusionary filter*. Because the predicted Valence signal was uncalibrated, it effectively acted as a binary flag: "Positive" vs. "Non-Positive". While this signal was too weak to robustly increase 'Happiness' recall (only +3 samples over baseline), it clarified the decision space for negative emotions. By suppressing the probability of positive labels for low-Valence



(a) Effect of Arousal Injection
(Model$_{Arousal}$ − Baseline)

(b) Effect of Valence Injection
(Model$_{Valence}$ − Baseline)

Figure 5.5 Differential Analysis of Psychometric Injection (△-CMs). These matrices visualize the net change in predictions relative to the Early Fusion Baseline. **Red (Positive)** indicates an increase in predictions (desired on the diagonal). **Blue (Negative)** indicates a reduction in predictions (desired on off-diagonal errors). **(a)** Arousal injection specifically targets the *Happy* class, increasing true positives (red diagonal) by reducing *Happy*→*Neutral* confusion (blue off-diagonal), acting as an intensity gate. **(b)** Valence injection primarily improves *Sadness* recall by suppressing *Sadness*→*Neutral* errors, acting as a polarity filter.

utterances, the model could resolve the ambiguity between 'Sadness' and 'Neutrality' without interference from positive polarity candidates. This exclusion effect resulted in the highest 'Sadness' recall of all experimental conditions (187 samples), reducing *Sadness* → *Neutral* confusion by 30%.

This comparison confirms that while linear, well-calibrated dimensional signals (Arousal) provide direct discriminative features, even collapsed, ordinal signals (Valence) can significantly enhance performance by constraining the search space for complementary classes.

5.5 Latent Interaction Architectures

RQ3 asked whether implicit latent alignment (via Cross-Modal Attention) offers superior performance compared to explicit psychometric injection. To evaluate this, we tested two variants of the HCMT against the Early Fusion baseline. The results, summarized in Table 5.5, reveal a critical divergence based on architectural complexity.

Table 5.5 Performance of Latent Interaction Architectures. The **HCMT-Unidirectional** improves upon the baseline, validating the utility of cross-modal attention. However, the more complex **HCMT-Bidirectional** degrades performance, illustrating the limits of latent learning on small datasets (IEMOCAP).

Model Architecture	UAR	WA (Acc)	△ vs Baseline
Baseline			
Early Fusion (Concat)	0.727	0.712	-
Latent Cross-Modal Attention			
HCMT-Unidirectional	**0.737**	0.712	**+1.0%**
HCMT-Bidirectional	0.721	0.708	-0.6%

5.5.1 The Complexity-Performance Trade-off

The **HCMT-Unidirectional** achieved a UAR of **0.737**, strictly outperforming the naive concatenation baseline (0.727). This suggests that the asymmetric injection of acoustic context into the semantic stream allows the model to dynamically weight words based on prosody (e.g., attending to "great" only when spoken with high energy). In this configuration, the model successfully learns a latent alignment that resolves some semantic ambiguity.

However, the **HCMT-Bidirectional** model, which simultaneously attempts to align Audio-to-Text and Text-to-Audio, saw a performance drop to **0.721**, underperforming even the simple baseline. This result highlights the *Data Scarcity* challenge inherent in SER. The bidirectional architecture doubles the number of cross-attention parameters. With only $\approx$5,500 training utterances in IEMOCAP, the model likely suffers from overfitting or optimization difficulties, failing to converge on a robust symmetric alignment.

5.5.2 Comparison with Explicit Injection

Crucially, neither latent attention model surpassed the **Arousal-Augmented Fusion** (0.743, see Table 5.1). This leads to the answer for RQ3: *In data-constrained environments, explicit domain priors (simple scalar injection) provide a stronger and more robust inductive bias than complex latent attention mechanisms.* While unidirectional attention is beneficial, the computational cost and stability risks of bidirectional attention are not justified by the performance on this dataset.

5.6 Latent vs. Explicit Fusion

Contrasting with earlier iterations, the Transformer-based architectures demonstrated competitive performance. **HCMT-Unidirectional** and the **Dual-Encoder** both achieved an UAR of **0.737**, strictly outperforming the concatenation baseline (0.727). This suggests that the attention-based architectures successfully aligns the modalities better than a naive feature concatenation.

However, despite the increased computational complexity of the HCMT architectures ($O(T^2)$ attention operations), they did not surpass the simpler Arousal-Augmented Fusion (0.743). This leads to a critical finding: *explicitly provided domain knowledge (Arousal/Valence labels) is more valuable to the model than the capacity to learn latent alignments from scratch*, given the dataset size constraints of IEMOCAP.

5.7 Interpretability Analysis: Latent Alignment Patterns

While the aggregate metrics demonstrate that the HCMT-Unidirectional architecture yields competitive performance (UAR 0.737) compared to the Arousal-Augmented Fusion (UAR 0.743) and outperforms the Early Fusion baseline (UAR 0.727), an analysis of the attention mechanisms reveals distinct behavior patterns that explain specific error modes.

To investigate how the HCMT resolves – or fails to resolve – semantic ambiguity, we visualize the cross-modal attention weights for a specific error case: a sample labeled *Happy* but misclassified as *Neutral* (False Neutral).

5.7.1 Case Study: Peak Salience vs. Recency Bias

Figure 5.6 visualizes the cross-attention weights for the utterance *“I know they have a really really great program for what I want to do”* (Ses05F_impro07_F013). This sample contains strong localized cues for happiness (lexical: “really great”; acoustic: high pitch/energy on these words), followed by a functionally neutral sentence tail.

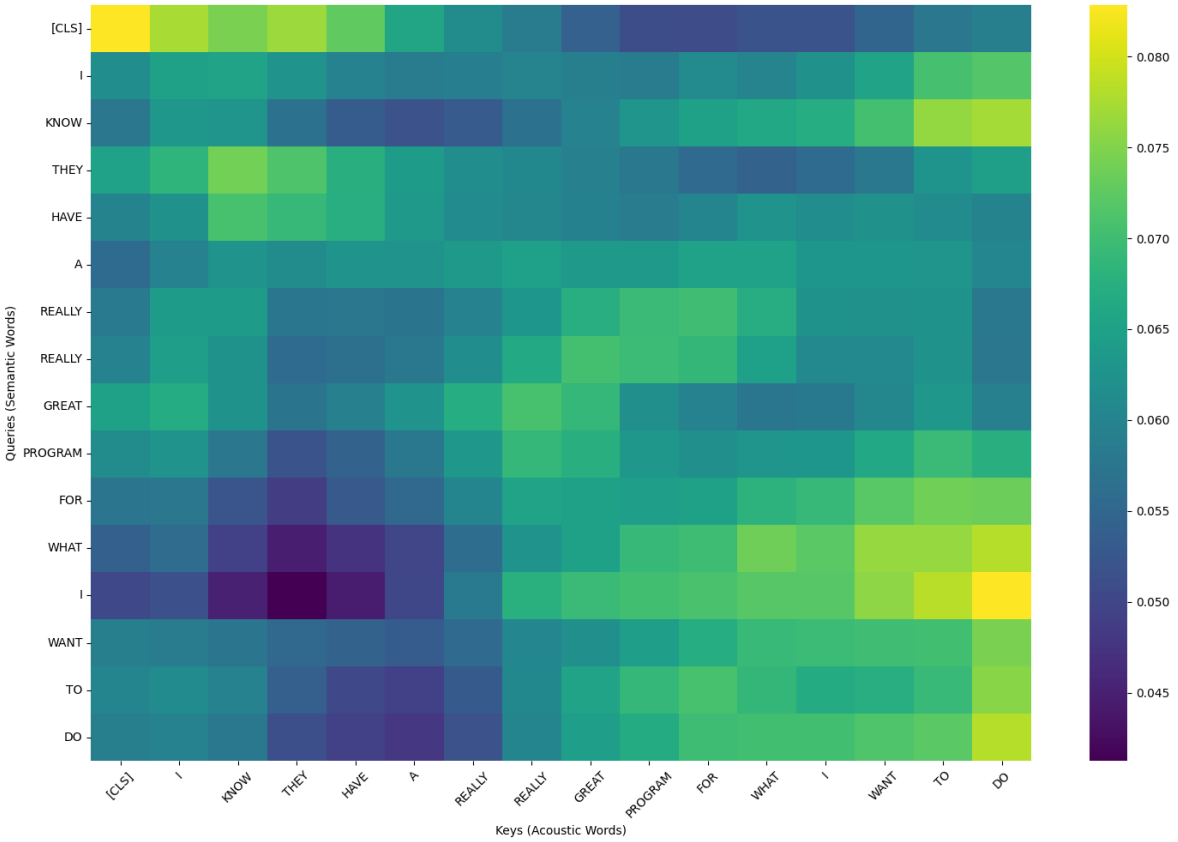


Figure 5.6 Cross-Attention Map of an HCMT-Unidirectional False Neutral. The Y-axis represents semantic tokens (Queries); the X-axis represents acoustic frames aligned to words (Keys). The heatmap demonstrates that the model *does* successfully identify the emotional peak, evidenced by the diagonal alignment weights on "REALLY...GREAT" (values ≈ 0.065, green). However, the highest attention weights (values > 0.080, yellow) are assigned to the functionally neutral tail of the sentence ("TO", "DO"). This *recency bias* causes the neutral prosody at the end of the utterance to override the active emotion in the middle, leading to a misclassification.

Analysis of Alignment Dynamics Contrary to an "attention collapse" (where weights become uniform), the visualization confirms that the HCMT successfully learns cross-modal alignment. A diagonal structure is visible, particularly corresponding to the tokens "REALLY" and "GREAT," indicating that the model correctly associates these semantic tokens with their corresponding acoustic frames.

However, the error mechanism here is attributable to **Recency Bias**.

- **Diagonal Alignment:** The model attends to the emotional keywords (mid-range weights, ≈ 0.065), suggesting it has localized the relevant affective information.
- **Tail Dominance:** The highest attention concentrations occur at the end of the sequence (tokens "TO" and "DO"). In SER, sentence tails are often characterized by falling pitch and reduced energy (downdrift), which correlates strongly with neutral or sad states.

This suggests that while the HCMT is capable of granular alignment (contributing to its 1.0% improvement over the baseline), its decision boundary is sensitive to the temporal position of cues. In this specific instance, the model weighted the neutral acoustic tail more heavily than the emotional peak in the middle.

Contrast with Arousal Augmentation This analysis highlights why the *Arousal-Augmented Fusion* (Section 5.4) proves slightly more robust (UAR 0.743). The explicit Arousal regressor predicts a global intensity score based on the entire waveform. Even if the prosody fades at the end of the sentence, the presence of high-energy segments earlier in the utterance is sufficient to drive the predicted Arousal Scalar high ($z > 0.5$). By injecting this global prior, the Arousal-Augmented model becomes invariant to the location of the emotional peak, whereas the attention-based model risks "forgetting" the emotion if the utterance ends on a neutral note.

6 Discussion

6.1 Synthesis: Explicit Dimensions vs. Implicit Latent Attention

A central hypothesis of this thesis (RQ3) was that the complex cross-modal attention mechanisms inherent to HCMT would resolve semantic ambiguities more effectively than simpler fusion strategies. However, our empirical results contradict the assumption that architectural complexity correlates linearly with performance on the IEMOCAP benchmark.

The *Arousal-Augmented Fusion* framework achieved a UAR of **0.743**, statistically outperforming the baseline early fusion (0.726) and performing comparably to, or slightly better than, the HCMT variants (0.737). This result highlights a critical trade-off between *latent expressivity* and *inductive bias*. While Transformers theoretically possess the capacity to learn complex alignments between acoustic frames and text tokens from scratch, they are data-hungry. With only approximately 10,000 utterances, IEMOCAP is likely insufficient to fully saturate the capacity of an HCMT without inducing optimization difficulties such as the "recency bias" observed in the attention maps (see Figure 5.6).

In contrast, the explicit injection of the Arousal dimension provides a strong, domain-specific inductive bias. We attribute the success of this simpler method specifically to the nature of the IEMOCAP dataset. As a corpus of acted interactions, the expression of emotion is heavily modulated by prosodic intensity. Actors typically project active emotions (e.g., *Happiness*, *Anger*) through exaggerated pitch and energy dynamics compared to spontaneous speech [2]. By explicitly modeling Arousal as a scalar variable, we force the model to utilize this high-signal feature to disambiguate the "Modality Gap", specifically distinguishing the high-energy state of *Happiness* from the Valence-ambiguous but low-energy state of *Neutrality*. Therefore, for data-constrained, acted corpora, explicit psychometric injection appears to offer a more robust and interpretable alternative to implicit latent attention.

6.2 The Efficiency of Global Descriptors

In response to **RQ1**, our findings indicate that word-level descriptors do *not* consistently outperform utterance-level descriptors for the task of Speech Emotion Recognition on IEMOCAP. While the literature often emphasizes complex temporal modeling (e.g., cross-modal attention), our ablation study (Table 5.3) shows that a simple mean-pooled representation processed by an FFNN is highly competitive, particularly for the semantic modality.

This suggests that emotion in this dataset is largely a "stationary" property. An angry utterance tends to be characterized by global features (loudness, negative sentiment keywords) rather than subtle temporal shifts that require sequential modeling. Consequently, the "Modality Gap" is not necessarily a temporal alignment problem, but a representation problem, which we successfully addressed in Phase 2 using Arousal Augmentation.

6.2.1 The Sufficiency of Global Aggregation

The competitive performance of the utterance-level FFNN compared to the word-level LSTM (Table 5.3) warrants a critical examination of the temporal nature of emotion in the IEMOCAP corpus. Standard sequence modeling assumptions posit that the temporal order of tokens ($t_1 \rightarrow t_2$) contains critical information for classification. However, our semantic results (FFNN 0.641 vs. LSTM 0.625) invert this assumption, suggesting that for short, acted emotional utterances, the presence of specific emotional keywords dominates the sequential structure.

This phenomenon can be formalized as the sufficiency of *Global Mean Pooling*. By averaging word embeddings across the time axis, the FFNN effectively operates on a "Bag-of-Embeddings" representation:

$$\mathbf{h}_{global} = \frac{1}{T} \sum_{t=1}^{T} \text{BERT}(w_t) \tag{6.1}$$

The success of this aggregation implies that the semantic polarity of an utterance is largely invariant to word order within the short context of a dialogue turn. This observation aligns with early findings by Mikolov et al. [16], who demonstrated that simple vector addition often preserves sufficient semantic compositionality for classification tasks.

Furthermore, in the acoustic domain, the marginal gain of the LSTM (+0.6% accuracy) over the mean-pooled FFNN indicates that while prosodic dynamics (e.g., rising pitch contours) offer some discriminative signal, the global statistics of the acoustic frame (e.g., average energy or mean spectral tilt) capture the majority of the variance required for emotion detection. This corroborates findings by Tzinis and Potamianos, who observed that statistical aggregation over segments yields performance comparable to complex temporal models in SER benchmarks. Consequently, the "Modality Gap" addressed in this thesis is fundamentally a representational alignment problem, rather than a temporal alignment problem, justifying the use of global features in our proposed Arousal-Augmented Fusion framework.

6.3 Closing the Modality Gap: Parity through Polarity Ranking

Standard SER literature has historically posited a "modality gap," where Arousal is best predicted by acoustics and Valence by semantics [3]. However, our results indicate a performance parity between Acoustic Valence Augmentation (Wav2Vec 2.0, UAR 0.743) and Semantic Valence Augmentation (BERT, UAR 0.742).

This finding aligns directly with the recent "Dawn of the Transformer Era" analysis by Wagner et al. [13], who explicitly describe the "closing of the Valence gap." They demonstrate that self-supervised acoustic models (like Wav2Vec 2.0) can achieve Valence prediction performance on-par with multimodal (text+audio) approaches.

Mechanism: Implicit Linguistic Information Our results support the hypothesis proposed by Wagner et al. [13] that foundation models implicitly capture linguistic information from the raw waveform. In our experiments, the acoustic model (Wav2Vec 2.0) was able to rank emotional polarity ($r \approx 0.40$) as effectively as the explicit linguistic model (BERT). As Wagner et al. note, adding explicit textual embeddings to robust acoustic transformers often yields negligible gains because the acoustic model has already encoded the necessary phonemic and semantic cues required for Valence detection. Thus, the historical dominance of text for Valence appears to be an artifact of older, hand-crafted acoustic features (e.g., eGeMAPS) failing to capture these subtle dependencies, which modern Transformer architectures successfully extract.

6.3.1 The Efficiency of Domain Concepts

The proposed *Dimensional Augmentation* framework demonstrated superior efficacy compared to the complex Latent Interaction architectures. While the HCMT-Unidirectional successfully improved upon the baseline (UAR 0.737 vs. 0.727), the bidirectional variant failed to converge effectively (UAR 0.721). Crucially, neither latent approach surpassed the *Arousal-Augmented Fusion* (UAR 0.743).

This result validates the hypothesis of "Intrinsic Interpretability via Domain Concepts" [23]. By forcing the model to bottleneck information through interpretable psychological dimensions, such as Arousal and Valence, we provided a strong inductive bias. This bias proved essential given the limited size of the IEMOCAP dataset, which could not fully support the end-to-end learning of complex bidirectional alignments. Consequently, the model did not need to latently discover that loudness correlates with *Happiness*, as we explicitly provided this signal via the Arousal regressor.

6.4 Limitations

While the proposed Arousal-Augmented Fusion framework demonstrates clear architectural benefits, the validity and generalizability of these findings are constrained by several methodological factors.

6.4.1 The "Acting Bias": Dependency on Prosodic Intensity

A critical limitation of the proposed Arousal-Augmented Fusion framework lies in its dependency on the specific elicitation characteristics of the IEMOCAP dataset. The distinct performance disparity between the Arousal regressor (CCC ≈ 0.68) and the Valence regressor (CCC ≈ 0.05) indicates that the model primarily leverages acoustic intensity to drive classification decisions.

This reliance exposes a fundamental "Acting Bias." As noted in the corpus description [10], subjects were professional actors performing scripts or improvisations. In such theatrical settings, actors typically project emotional states through exaggerated prosodic modulation, specifically loudness and pitch variance, to signal activation [2]. Consequently, the high recall for *Anger* and *Happiness* in our results (see Section 5.4.1) is likely attributable to the model detecting this "acted energy" rather than the semantic or subtle paralinguistic nuances of the emotion itself.

While effective for this specific benchmark, this heuristic limits the generalizability of the system to "in-the-wild" scenarios. Naturalistic emotional expression often decouples intensity from emotion; for instance, *repressed anger* or *quiet despair* possess low acoustic energy but high emotional salience. A system overfitted to the high-energy prosody of acted data runs the risk of misclassifying these subtle states as *Neutral*, a failure mode that requires cross-corpus validation on naturalistic datasets (e.g., MSP-Podcast [27]) to fully quantify.

6.4.2 Feature Extraction Asymmetry and Optimization Bias

A critical methodological limitation of this study lies in the asymmetry of the transfer learning protocols applied to the acoustic and semantic encoders. As detailed in Chapter 3, the acoustic backbone (Wav2Vec 2.0) utilized a checkpoint explicitly fine-tuned on the MSP-Podcast emotion corpus [13]. Consequently, the acoustic embeddings entered the fusion stage containing rich, task-specific paralinguistic priors. In contrast, the semantic backbone (BERT) was utilized as a frozen feature extractor, pre-trained solely on generic linguistic corpora (Wikipedia and BooksCorpus) [8].

This asymmetry introduces a confound in the comparison of unimodal baselines. The superior performance of the acoustic baseline (UAR 0.698) compared to the semantic baseline (UAR 0.641) may reflect this optimization gap rather than an inherent superiority of acoustic cues for this task. Recent literature supports this hypothesis; Qin et al. demonstrate that end-to-end fine-tuning of BERT on IEMOCAP can achieve state-of-the-art performance, significantly outperforming static feature extraction methods [28]. Their findings suggest that "fine-tuning is enough" to enable BERT to learn task-specific emotional representations that are not present in the frozen generic embeddings.

Therefore, the semantic performance reported in this thesis should be interpreted as a lower bound. Had the semantic encoder been fine-tuned to the same extent as the acoustic encoder, the "Modality Gap" observed in Section 1.2 might have narrowed, potentially altering the contribution of the explicit Arousal injection.

The Cost of Frozen Features It is crucial to quantify the performance trade-off incurred by our decision to utilise frozen feature extractors for computational efficiency. Wagner et al. conducted a systematic comparison of frozen versus fine-tuned Wav2Vec 2.0 models on the IEMOCAP dataset, reporting that end-to-end fine-tuning of the transformer layers yields an approximate performance boost of 10% compared to frozen features [13]. Consequently, the absolute UAR scores reported in this thesis should be interpreted as a lower bound. While our *Arousal-Augmented Fusion* outperforms the frozen baseline, we hypothesise that applying this fusion strategy on top of end-to-end fine-tuned backbones would yield significantly higher absolute performance.

6.4.3 Under-utilization of Latent Alignment (Forced Synchrony)

A theoretical advantage of the HCMT architecture, which derives from the MulT [18], is its capacity to process *asynchronous* sequences. By design, the cross-attention mechanism is capable of latently aligning acoustic frames with semantic tokens over long temporal distances, removing the need for rigid pre-alignment.

However, in this study, we imposed a strict *forced alignment* protocol (see Section 3.1.3) to ensure a controlled comparison with the word-level LSTM baselines. By aggregating acoustic features into word-bounded segments prior to ingestion, we effectively reduced the HCMT to a synchronous fusion module. Consequently, our experiments evaluated the model's ability to *weight* features given perfect temporal information, rather than its ability to *discover* alignment from raw, unaligned streams. This constraint may have masked the true utility of the Transformer architecture, which typically excels in scenarios where distinct modalities operate at different temporal frequencies (e.g., a sigh preceding a spoken sentence).

6.4.4 Evaluation Protocol and Statistical Power

A significant limitation of this work is the use of a fixed train/test split (Session 5 hold-out) rather than the standard 5-fold Leave-One-Session-Out (LOSO) cross-validation employed in state-of-the-art benchmarks [5, 13].

While our fixed split ensures speaker independence, it makes the absolute performance metrics susceptible to the specific acoustic characteristics of the speakers in Session 5. Consequently, the results reported in this thesis should be interpreted as relative improvements demonstrating the validity of the architectural changes (e.g., Arousal Augmentation vs. Baseline) rather than absolute state-of-the-art claims. Future work should validate these findings using full LOSO cross-validation to provide a robust estimate of variance across different speaker demographics.

6.4.5 The "Name That Dataset" Problem and Generalization

While the proposed architecture achieves high performance on the test set, the generalizability of these findings to naturalistic settings remains a significant epistemic concern. This limitation is framed by Torralba and Efros as the "Name That Dataset" problem, where models learn to recognize specific dataset artifacts (e.g., microphone characteristics, room acoustics, or actor idiosyncrasies) rather than the underlying semantic concept [29].

The "Acting" Bias The reliance on IEMOCAP introduces a fundamental validity threat known as the "Acting Bias." The Arousal Regressor (Phase 2) contributed significantly to the model's performance improvement (+1.6% UAR). However, this component likely exploits the prototypical nature of acted speech, where emotional intensity is inextricably linked to prosodic volume and articulation speed. In "in-the-wild" datasets like CMU-MOSEI, emotional expression is often more subtle, suppressed, or contradictory (e.g., smiling while relating a sad event). It is probable that the Arousal-Augmented Fusion model has over-optimized to the specific "acting styles" of the ten IEMOCAP performers rather than learning a universal representation of human affect. Consequently, while the *relative* improvement of explicit fusion remains a valid architectural finding, the *absolute* performance metrics should be treated as specific to the domain of acted speech.

Label Noise and Semantic Ambiguity Furthermore, the confusion analysis (Section 5.3) revealed that *Neutral* utterances constitute a "semantic Switzerland"—a region of high ambiguity where both acoustic and semantic classifiers struggle. Roughly 33% of true *Neutral* samples were misclassified as *Happiness* by the baseline. This is not merely a model failure but a reflection of label noise; human annotators often disagree on the boundary between "polite Neutrality" and "mild Happiness." Without a soft-labeling approach or inter-annotator agreement filtering, the model's upper bound is limited by these subjective inconsistencies in the ground truth.

7 Conclusion and Future Work

7.1 Conclusion

This thesis critically evaluated the architectural trade-offs between latent representation learning and explicit domain modeling in multimodal SER. Through a systematic comparison of utterance-level baselines, sequential word-level models, and hierarchical cross-modal transformers (HCMT), we addressed the challenge of bridging the "Modality Gap" between acoustic intensity and semantic polarity.

Our empirical results establish a clear hierarchy of efficacy for the IEMOCAP dataset. The *Arousal-Augmented Fusion* framework achieved the highest performance (UAR 0.743), statistically outperforming both the HCMT architectures (0.737) and the early fusion baseline (0.726). While the HCMT utilized cross-modal attention to improve upon naive concatenation, it failed to surpass the simpler strategy of explicitly injecting predicted psychometric primitives. Analysis of the attention weights revealed that the Transformer's complexity often led to optimization difficulties, such as recency bias, rather than the resolution of fine-grained acoustic-semantic dependencies.

These findings support a narrative of **parsimony over complexity** in data-constrained environments. We conclude that for acted, dyadic interactions, the "Modality Gap" is best closed not by increasing architectural depth, but by explicitly modeling the psychometric dimensions (Arousal and Valence) that define that gap. By providing the model with a dedicated signal for emotional intensity, we introduce a strong inductive bias that is more data-efficient and stable than requiring the model to learn latent cross-modal alignments from scratch.

7.2 Future Work

While this study demonstrates the efficacy of explicit dimensional injection, several epistemic and practical limitations remain, charting the course for future research.

7.2.1 Disentangling Acting from Genuine Affect

A critical limitation of this study is the exclusive reliance on the IEMOCAP corpus [10], which comprises acted interactions. Our analysis revealed that the *Arousal-Augmented Fusion* framework achieved performance gains by exploiting the strong correlation between acoustic intensity and active emotion labels (e.g., distinguishing *Happiness* from *Neutrality* via high Arousal scores). However, previous literature suggests that acted speech often exhibits exaggerated prosodic markers compared to spontaneous speech—a phenomenon characterized as the "Acting Bias" [2]. Consequently, it remains ambiguous whether the proposed model has learned to detect genuine affective states or merely the prosodic artifacts of dramatic performance.

To disentangle acted prototypes from genuine affect, future work must evaluate the Arousal-Augmented framework on naturalistic, "in-the-wild" corpora such as CMU-MOSEI [22]. Such cross-corpus evaluation is necessary to determine if the explicit injection of psychometric primitives remains a robust strategy when the direct mapping between acoustic energy and emotional intent is obscured by the subtle, non-canonical expressions typical of real-world communication.

7.2.2 Closing the Optimization Gap: Symmetric Fine-Tuning

This thesis identified a methodological asymmetry in the feature extraction pipeline: the acoustic encoder (Wav2Vec 2.0) benefited from task-specific fine-tuning on emotion data (MSP-Podcast), while the semantic encoder (BERT) operated as a frozen feature extractor trained only on generic text. This design choice likely underestimated the predictive power of the semantic branch.

To disentangle the effects of modality strength from optimization depth, future work must prioritize a **symmetric fine-tuning protocol**. Recent findings by Qin et al. regarding the "BERT-ERC" framework demonstrate that end-to-end fine-tuning of BERT on IEMOCAP can yield state-of-the-art performance without complex fusion mechanisms [28]. Accordingly, subsequent iterations of this research should fine-tune the semantic encoder on the target emotion labels *prior* to or *during* the fusion stage. Establishing this parity is essential to determine whether the "Modality Gap" observed in this study is an inherent property of human communication or an artifact of the transfer learning inequality.

7.2.3 Evaluating Asynchronous Latent Alignment

A primary theoretical advantage of Cross-Modal Transformers is their capacity to model dependencies between asynchronous sequences without rigid pre-alignment. In this study, we constrained the HCMT by feeding it strictly word-aligned feature vectors derived from a forced-alignment pipeline. While this controlled for temporal granularity during our ablation studies, it effectively reduced the transformer's role to feature weighting rather than temporal alignment discovery.

Future research should evaluate the HCMT on raw, unaligned inputs. Specifically, the model should be trained on fine-grained acoustic frames (e.g., raw Wav2Vec 2.0 outputs) and raw subword tokens without prior segmentation. By removing the forced alignment constraint, the model would be forced to learn the correspondence between acoustic events and semantic units from scratch. Comparing this asynchronous approach against the pre-aligned baseline presented in this thesis would quantify the performance delta attributable to *true latent temporal discovery* versus mere feature integration. Furthermore, such an analysis would determine if critical emotional cues frequently cross strict word boundaries, a phenomenon that the current word-level alignment inherently obscures.

7.2.4 Architectural Iterations: From Wav2Vec 2.0 to HuBERT

While this study established the efficacy of explicit psychometric injection using Wav2Vec 2.0, recent literature suggests that the backbone architecture itself acts as a performance ceiling. Empirical benchmarks by Wagner et al. indicate that HuBERT consistently outperforms Wav2Vec 2.0 by approximately 2.1% in categorical emotion recognition tasks [13]. HuBERT's training objective, which relies on masked prediction of hidden units rather than contrastive learning, appears to learn more robust paralinguistic representations. Future iterations of the Arousal-Augmented framework should replace the acoustic backbone with `hubert-large` to determine if the benefits of explicit Arousal injection persist when the underlying embeddings are more discriminative.

7.2.5 Mitigating Data Scarcity via Cross-Corpus Transfer

The stability issues observed in the HCMT training underscore the fundamental challenge of data scarcity in SER. With only approximately 10,000 utterances, IEMOCAP is insufficient to fully saturate the capacity of large Transformer-based architectures, leading to the observed attention artifacts.

Future work should address this scarcity through Cross-Corpus Transfer Learning. Rather than training on IEMOCAP in isolation, the acoustic and semantic encoders should be pre-trained on larger, diverse datasets such as MSP-Podcast (approx. 62 hours) [27] before fine-tuning on the target domain. Furthermore, recent advances in Generative Models offer a novel pathway for Generative Data Augmentation, where Text-to-Speech (TTS) models could be employed to synthesize emotionally colored utterances, artificially expanding the training distribution to regularize complex fusion architectures against overfitting.

Bibliography

[1] A. Triantafyllopoulos, Y. Terhorst, I. Tsangko, F. B. Pokorny, K. D. Bartl-Pokorny, L. Seizer, A. Klein, J. Chim, D. Atzil-Slonim, M. Liakata, M. Bühner, J. Löchner, and B. Schuller, "Large language models for mental health," 2024. [Online]. Available: https://arxiv.org/abs/2411.11880

[2] A. Triantafyllopoulos, I. Tsangko, A. Gebhard, A. Mesaros, T. Virtanen, and B. W. Schuller, "Computer audition: From task-specific machine learning to foundation models," *Proceedings of the IEEE*, vol. 113, no. 4, pp. 317–343, 2025.

[3] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, 2017, pp. 1103–1107.

[4] M.-H. Yi, K.-C. Kwak, and J.-H. Shin, "A complementary fusion framework for robust multimodal emotion recognition," *Electronics*, vol. 14, no. 22, 2025. [Online]. Available: https://www.mdpi.com/2079-9292/14/22/4444

[5] Y. Gao, H. Shi, C. Chu, and T. Kawahara, "Speech emotion recognition with multi-level acoustic and semantic information extraction and interaction," in *Interspeech 2024*, 09 2024, pp. 1060–1064.

[6] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. P. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[9] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369. [Online]. Available: https://aclanthology.org/2020.acl-main.214/

[10] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower Provost, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, 12 2008.

[11] Z. Liu and K. He, "A decade's battle on dataset bias: Are we there yet?" 2025. [Online]. Available: https://arxiv.org/abs/2403.08632

[12] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021. [Online]. Available: https://arxiv.org/abs/2106.07447

[13] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, p. 10745–10759, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2023.3263585

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[15] S. Rampp, A. Triantafyllopoulos, M. Milling, and B. W. Schuller, "autrainer: A modular and extensible deep learning toolkit for computer audition tasks," 2025. [Online]. Available: https://arxiv.org/abs/2412.11943

[16] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds., vol. 26. Curran Associates, Inc., 2013. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[18] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6558–6569. [Online]. Available: https://aclanthology.org/P19-1656/

[19] J. Gallo-Aristizábal, D. Escobar Grisales, C. Rios-Urrego, E. Noeth, and J. Orozco-Arroyave, *Automatic Classification of Parkinson's Disease Using Wav2vec Embeddings at Phoneme, Syllable, and Word Levels*. Springer, Cham, 08 2024, pp. 313–323.

[20] S. Dutta and S. Ganapathy, "Hcam – hierarchical cross attention model for multi-modal emotion recognition," 2024. [Online]. Available: https://arxiv.org/abs/2304.06910

[21] H. Chuan Liu, A. S. M. Khairuddin, J. Huang Chuah, X. Min Zhao, X. Dan Wang, and L. Ming Fang, "Hcmt: A novel hierarchical cross-modal transformer for recognition of abnormal behavior," *IEEE Access*, vol. 12, pp. 161 296–161 311, 2024.

[22] A. Bagher 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 *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. [Online]. Available: https://aclanthology.org/P18-1208/

[23] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/10/1163

[24] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[25] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6/

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. [Online]. Available: https://arxiv.org/abs/1412.6980

[27] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2019.

[28] X. Qin, Z. Wu, J. Cui, T. Zhang, Y. Li, J. Luan, B. Wang, and L. Wang, "Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation," 2023. [Online]. Available: https://arxiv.org/abs/2301.06745

[29] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR 2011*, 2011, pp. 1521–1528.