SEER: Label-Structured Modality Routing for Multimodal Sentiment Analysis and Intent Recognition

Anonymous authors
Paper under double-blind review

Abstract

Multimodal sentiment analysis and intent recognition require models to combine textual, acoustic, and visual evidence whose reliability varies across utterances. Although adaptive fusion can address this variability by assigning sample-specific modality weights, many existing routing mechanisms estimate confidence from raw feature statistics, generic similarity measures, or prototype assignments that are only indirectly related to the downstream label structure. This can make routing sensitive to modality style or feature magnitude rather than to the evidence most relevant for sentiment or intent prediction. To study this issue, we introduce a staged routing framework. First, Emotion-Aware Modality Calibration (EAMC) serves as an encoded-space routing baseline that estimates modality reliability after semantic encoding while keeping the backbone and weighted-sum fusion rule fixed. Building on this baseline, we propose Structured Evidence Estimation and Routing (SEER), which incorporates label structure into the representation space used for confidence estimation. SEER-L0 adds label-aware contrastive supervision to organize modality representations according to task labels, while SEER-L1 estimates modality confidence by matching modality-adapted representations to shared label-structured anchors. We also evaluate SEER-L2, a prototype-guided temporal evidence extraction extension. Experiments on aligned CMU-MOSI, aligned CMU-MOSEI, and MIntRec under a multi-run evaluation protocol show that SEER-L1 provides the most consistent improvement over EAMC on the primary metrics, namely binary F1 for sentiment analysis and Weighted F1 for intent recognition. In contrast, SEER-L2 does not improve performance in the current setting. Overall, our findings support the claim that adaptive multimodal routing is more effective when modality confidence is estimated in a label-grounded semantic space.

1 Introduction

Multimodal sentiment analysis and intent recognition aim to infer affective or communicative states from textual, acoustic, and visual signals. Over the past several years, this area has progressed from explicit fusion operators toward richer multimodal interaction and representation-learning frameworks, including tensorized fusion, low-rank factorization, cross-modal attention, shared-private decomposition, and self-supervised or information-theoretic objectives (Zadeh et al., 2017; Liu et al., 2018; Tsai et al., 2019; Hazarika et al., 2020; Yu et al., 2021; Han et al., 2021; Li et al., 2023). Despite these advances, a persistent challenge is that the informativeness of each modality varies across utterances. Lexical content may be sufficient in some cases, while prosody, facial behavior, or cross-modal disagreement may be more informative in others, as in sarcasm or masked affect. This variability makes a single static fusion rule poorly matched to the structure of the task.

This observation has motivated adaptive fusion and expert-routing methods that assign sample-specific modality weights or routing decisions rather than applying the same fusion mechanism to every utterance (Gao et al., 2024; Fang et al., 2025; Chen et al., 2025). However, in representative adaptive-routing approaches, the routing signal is typically derived from feature activations, uncertainty estimates, expert outputs, or cross-modal agreement patterns rather than from a representation explicitly organized around

the downstream label structure. Routing may therefore depend on raw feature statistics, generic similarity scores, or learned prototypes-reference vectors used for similarity-based matching-whose organization is not directly constrained by sentiment or intent labels. Such signals can capture magnitude, modality style, uncertainty, or cross-modal agreement, but they do not directly estimate which modality provides the most label-relevant evidence for the current prediction. As a result, adaptive fusion can remain difficult to supervise and interpret when modality reliability depends on task-specific semantics.

We address this issue by studying how label information can be incorporated into the routing space. Three limitations motivate the proposed formulation. First, routing decisions made before semantic encoding cannot assess modality relevance using task-level representations. Second, such prototype-based routing is limited when the prototypes are learned without explicit alignment to sentiment or intent structure. Third, standard instance-level contrastive objectives such as NT-Xent can distort the routing geometry by treating semantically related samples as negatives, even when they share the same class label or occupy nearby positions on an ordinal sentiment scale (Khosla et al., 2020). These observations suggest that adaptive routing should be learned in a representation space that is both semantically encoded and structured by the downstream labels.

To isolate these factors, we first define Emotion-Aware Modality Calibration (EAMC), an encoded-space routing baseline that moves modality reliability estimation from raw input features to encoded representations while preserving the shared multimodal backbone and weighted-sum fusion rule. EAMC addresses a basic question: how much is gained by estimating modality reliability after semantic encoding rather than before it? Building on this baseline, we introduce Structured Evidence Estimation and Routing (SEER), a framework that uses label information to structure modality confidence estimation. SEER-L0 adds label-aware contrastive supervision to the encoded routing space. SEER-L1 then replaces unconstrained prototypes with a shared-private anchor structure: modality-specific private prototypes adapt each modality into a style-specific subspace, while shared anchors convert the adapted representation into a label-structured confidence signal. We also evaluate prototype-attentive temporal evidence extraction as an optional extension, but treat it separately from the main routing comparison.

We evaluate this sequence of models on CMU-MOSI and CMU-MOSEI for sentiment regression and on MIntRec for intent classification (Zadeh et al., 2016; Bagher Zadeh et al., 2018; Zhang et al., 2022). Across these benchmarks, the clearest improvements are obtained with SEER-L1. Relative to the encoded-space routing baseline, SEER-L1 improves the primary F1-style metric on the aligned sentiment benchmarks and obtains the highest Weighted F1 on MIntRec among the completed variants. In contrast, prototype-attentive temporal evidence extraction does not improve performance in the present setting. These results indicate that, for the evaluated benchmarks, improving the space in which modality confidence is estimated is more useful than adding temporal pooling complexity. In this implementation, the encoded-space routing family also uses a substantially smaller non-BERT routing and fusion stack than a raw-feature routing design.

Our contributions are as follows:

- We study adaptive multimodal routing as a problem of task-structured confidence estimation, and identify three limitations of common routing pipelines: routing before semantic encoding, prototype assignments without explicit label alignment, and contrastive objectives that can introduce false negatives among semantically related samples.
- We introduce EAMC as an encoded-space routing baseline and SEER as an extension that incorporates label structure into the routing representation and confidence-estimation mechanism.
- We evaluate the resulting model sequence on three benchmarks under a multi-run protocol, reporting both predictive performance and routing complexity through total and non-BERT parameter counts.
- We provide ablations and supplementary analyses showing that the most reliable improvement comes from shared-private label-structured routing, while prototype-attentive temporal evidence extraction does not improve performance in the current setting.

The remainder of the paper is organized as follows. Section 2 reviews related work, Section 3 presents the proposed method, Section 4 reports the main experiments, Section 5 discusses the findings and limitations, and Section 6 concludes. Additional results are provided in the appendix.

2 Related Work

2.1 Multimodal sentiment analysis and intent recognition

Multimodal sentiment analysis has developed from explicit fusion operators toward models that capture cross-modal interaction and modality-specific structure. Early methods such as TFN and LMF introduced structured fusion through tensorized or factorized interactions over text, audio, and visual streams (Zadeh et al., 2017; Liu et al., 2018). Subsequent work modeled cross-modal dependencies and modality-specific representations through cross-modal attention in MulT, shared-private decomposition in MISA, self-supervised multi-task learning in Self-MM, hierarchical mutual-information objectives in MMIM, and decoupling or distillation strategies for affective recognition (Tsai et al., 2019; Hazarika et al., 2020; Yu et al., 2021; Han et al., 2021; Li et al., 2023). CMU-MOSI and CMU-MOSEI have become widely used benchmarks for multimodal sentiment analysis (Zadeh et al., 2016; Bagher Zadeh et al., 2018). For intent recognition, MIntRec extends this setting beyond sentiment regression by providing a benchmark with text, audio, and video streams paired with intent labels (Zhang et al., 2022). Recent survey work provides a broader overview of multimodal intent recognition and its open challenges (Zhu et al., 2025). Although these methods have improved multimodal representation learning and fusion, many still use a fixed fusion procedure once modality features have been computed. This can be restrictive because the relative usefulness of text, audio, and vision often changes across utterances. The present work therefore focuses on sample-specific modality weighting, with particular attention to how modality reliability is estimated before fusion.

2.2 Adaptive fusion and mixture-of-experts for multimodal learning

Adaptive routing is closely related to the broader mixture-of-experts literature, which began with adaptive local experts and later developed into sparse conditional computation in modern deep neural networks (Jacobs et al., 1991; Shazeer et al., 2017; Fedus et al., 2022). In multimodal learning, expert-based models have been used to allocate computation across inputs or modalities and to encourage specialization. LIMoE showed that sparse experts can learn modality-sensitive behavior in large-scale multimodal contrastive learning (Mustafa et al., 2022). More recently, expert-routing ideas have been brought into multimodal sentiment analysis more directly. EUAR introduces uncertainty-aware routing, EMOE uses modality-specific dynamic experts for multimodal emotion and intent recognition, and MMA studies mixture-of-adapter fusion for sentiment analysis (Gao et al., 2024; Fang et al., 2025; Chen et al., 2025).

SEER belongs to this adaptive-fusion line, especially in relation to EMOE. The distinction lies in the routing signal rather than in the use of dynamic weighting itself. Existing methods often estimate routing weights from raw features, uncertainty estimates, expert activations, or cross-modal agreement patterns. In contrast, SEER examines whether modality confidence should be computed after semantic encoding and aligned with the downstream label structure. This shifts the emphasis from the fusion rule alone to the representation space in which modality reliability is measured.

2.3 Contrastive supervision and label-structured representation learning

Contrastive learning has become a standard tool for shaping representation geometry. SimCLR popularized instance-discrimination objectives such as NT-Xent/InfoNCE (Chen et al., 2020), while supervised contrastive learning showed that label information can be used directly to define semantically meaningful neighborhoods (Khosla et al., 2020). In multimodal learning, contrastive objectives have also been used to align representations across modalities, as in CLIP and later multimodal contrastive models such as LIMoE (Mustafa et al., 2022; Radford et al., 2021).

In this work, contrastive supervision serves a narrower purpose than general multimodal pretraining. We use label-aware contrastive structure to organize the modality representations used by the router. This

distinction is important for sentiment regression, where nearby sentiment values should not be treated as hard negatives. The contrastive term is therefore used to improve the geometry available to routing and prototype matching, rather than as a standalone representation-learning objective.

2.4 Prototype-based and shared-private modeling

Prototype-based learning organizes an embedding space around learned reference vectors, often used as class or cluster representatives. Prototypical Networks are a common example, using class prototypes for metric-based classification (Snell et al., 2017). Shared-private modeling has also been used in multimodal learning to separate task-relevant common structure from modality-specific variation (Hazarika et al., 2020). These ideas are useful for modality routing because shared components can represent task-level structure, while private components can preserve modality-specific expression patterns.

Our work connects these ideas at the routing stage. Instead of using prototypes only for final classification, or applying shared-private decomposition only within the encoder, SEER estimates modality confidence through a shared-private prototype structure. Modality-specific private prototypes adapt each modality into a style-specific space, and shared anchors evaluate the adapted representation with respect to the label structure. This design combines prototype matching and shared-private modeling to estimate modality reliability in a task-structured representation space.

3 Method

Adaptive multimodal routing depends strongly on the representation space used to estimate modality reliability. We therefore present the method as a sequence of post-encoding routing variants rather than as a single composite architecture. We begin with **Emotion-Aware Modality Calibration** (EAMC), which moves reliability estimation from raw input features to encoded modality representations while preserving the shared backbone and weighted-sum fusion rule. We then introduce **SEER-L0**, which adds label-aware supervision to the encoded routing space, and **SEER-L1**, which estimates modality confidence by matching modality-specific adapted representations to shared label-structured anchors. We also evaluate **SEER-L2**, an extension that replaces pooled modality representations with prototype-guided temporal evidence. Figure 1 summarizes these variants. Relation-aware expert fusion and robustness-oriented modality dropout were also explored, but they introduce additional assumptions beyond the routing mechanisms studied in the main paper; we therefore report them as supplementary extensions in Appendix A.

3.1 Overview and shared backbone

Let $\mathcal{D} = \{(X_i^t, X_i^a, X_i^v, y_i)\}_{i=1}^N$ denote a multimodal dataset, where X_i^t, X_i^a, and X_i^v are the text, audio, and visual streams of sample i, and y_i is either a sentiment target or an intent label. Let $\mathcal{M} = \{t, a, v\}$ denote the modality set. Following the backbone used across all variants, each modality is encoded into a temporal sequence

$$H_i^m = \mathrm{Enc}_m(X_i^m) \in \mathbb{R}^{T_m \times d}, \qquad m \in \mathcal{M},$$

where T_m denotes the modality-specific sequence length after preprocessing and d is the hidden dimension. In the default path, the sequence is summarized by its final hidden state,

$$h_i^m = H_i^m[T_m] \in \mathbb{R}^d.$$

Unimodal heads predict from each h_i^m, while the fused head predicts from a multimodal representation u_i. The proposed variants leave this encoder family unchanged and modify only the mapping from modality representations $\{h_i^m\}_{m \in \mathcal{M}}$ to confidence scores and routing weights.

3.2 Emotion-Aware Modality Calibration (EAMC)

EAMC provides the encoded-space routing baseline used in our comparisons. It keeps the shared backbone and weighted-sum fusion rule fixed, but estimates modality reliability from encoded representations rather

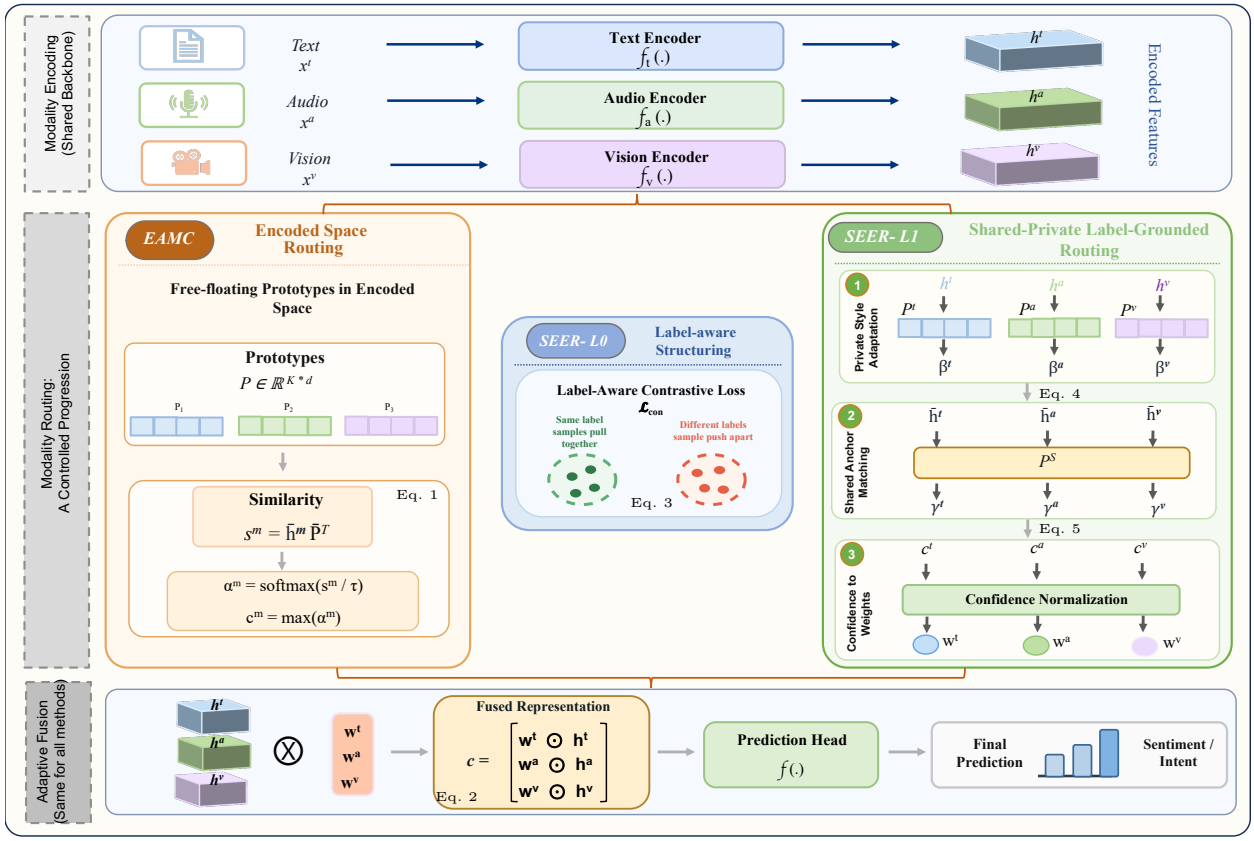

Figure 1: Overview of the routing variants. Text, audio, and vision are encoded by a shared backbone into modality representations h^t, h^a, h^v. EAMC estimates modality reliability in encoded feature space using a learnable prototype bank. SEER-L0 keeps this router unchanged and adds label-aware contrastive supervision to the encoded representations. SEER-L1 replaces the unconstrained prototype bank with a shared-private structure: modality-specific private prototypes produce style-adapted representations, shared anchors map these representations to modality confidence scores, and the resulting scores are normalized into routing weights for weighted fusion. Equation references indicate the corresponding stages of the computation.

than from raw input features. Let $P \in \mathbb{R}^{K \times d}$ be a learnable bank of K prototypes, where each prototype is a reference vector used for similarity-based matching. For each modality representation h_i^m, EAMC computes a soft assignment over prototypes using normalized similarity,

$$\alpha_i^m = \mathrm{softmax}\left(\frac{\bar{h}_i^m \bar{P}^\top}{\tau}\right), \qquad c_i^m = \max_k \alpha_{i,k}^m, \tag{1}$$

where $\bar{h}_i^m$ denotes the ℓ_2-normalized modality representation, $\bar{P}$ denotes the row-wise ℓ_2-normalized prototype bank, and τ is the routing temperature. The modality confidences are normalized across modalities to obtain

$$w_i^m = \frac{\exp(c_i^m/\tau)}{\sum_{m' \in \mathcal{M}} \exp(c_i^{m'}/\tau)}, \qquad m \in \mathcal{M},$$

The fused representation is then

$$u_i = \sum_{m \in \mathcal{M}} w_i^m h_i^m. \tag{2}$$

EAMC isolates the effect of moving reliability estimation into the encoded feature space. However, its prototypes are not constrained by the downstream sentiment or intent labels. A high prototype assignment can therefore indicate a confident match to the learned prototype bank without necessarily indicating reliable evidence for the target label. SEER addresses this limitation by introducing label structure into the routing representation and confidence-estimation process.

3.3 SEER: label-grounded routing

SEER extends EAMC by separating two aspects of the routing problem. The first is whether the encoded routing space should be organized using label-aware supervision. The second is whether modality confidence should be estimated with respect to anchors that represent the downstream label structure, rather than with respect to an unconstrained prototype bank. SEER-L0 addresses the first aspect, while SEER-L1 addresses the second.

Label-aware contrastive structuring (SEER-L0): SEER-L0 keeps the EAMC router unchanged and augments training with a label-aware contrastive term. Let $z_i^m = g_m(h_i^m)$ denote the output of a modality-specific projection head for modality m. The contrastive loss is applied independently to each modality branch and then averaged across modalities. For brevity, we write the loss for a single branch:

$$\mathcal{L}_{\text{con}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j\neq i}\bar{\omega}_{ij}\log\frac{\exp(\text{sim}(z_i, z_j)/\tau_c)}{\sum_{k\neq i}\exp(\text{sim}(z_i, z_k)/\tau_c)}. \tag{3}$$

Here $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ_c is the contrastive temperature, and $\bar{\omega}_{ij}$ determines the positive weight assigned to sample j for anchor i. For classification, $\bar{\omega}_{ij}$ is uniform over samples with the same label and zero otherwise. For regression, we use soft label-distance weights

$$\omega_{ij} = \exp\left(-\frac{|y_i - y_j|}{\sigma}\right), \qquad \bar{\omega}_{ij} = \frac{\omega_{ij}}{\sum_{k\neq i}\omega_{ik}},$$

where $\sigma > 0$ is a bandwidth parameter controlling how quickly the weight decreases with label distance. Thus, closer sentiment values receive larger positive weights. SEER-L0 modifies the training objective without changing the router itself.

Shared-private label-grounded prototype routing (SEER-L1): SEER-L1 replaces the EAMC prototype bank with a shared-private structure that separates modality-specific variation from task-level label structure. Here, private refers to modality-specific prototype banks, while shared refers to anchors used across all modalities to represent the downstream label structure. Let $P_s \in \mathbb{R}^{K_s\times d}$ denote a bank of shared anchors, where each anchor is a reference vector associated with the downstream label structure, and let

$$P^m_{\text{priv}} \in \mathbb{R}^{K_p\times d}, \qquad m \in \mathcal{M},$$

denote modality-specific private prototype banks. For classification, the shared bank contains one anchor per class; for regression, it contains ordered anchors spanning the sentiment label range. For each modality, we first compute a private style assignment and form a style-adapted representation,

$$\beta_i^m = \text{softmax}\left(\frac{\bar{h}_i^m(\bar{P}^m_{\text{priv}})^\top}{\tau}\right), \qquad \tilde{h}_i^m = \beta_i^m P^m_{\text{priv}}, \tag{4}$$

which is then matched to the shared anchors:

$$\begin{aligned}\gamma_i^m &= \text{softmax}\left(\frac{\bar{\tilde{h}}_i^m \bar{P}_s^\top}{\tau}\right), \qquad c_i^m = \max_k \gamma_{i,k}^m,\\ w_i^m &= \frac{\exp(c_i^m/\tau)}{\sum_{m'\in\mathcal{M}}\exp(c_i^{m'}/\tau)}, \qquad m \in \mathcal{M}.\end{aligned} \tag{5}$$

The final multimodal representation is then computed using the same weighted-sum fusion rule in Eq. 2.

This routing mechanism changes the source of the confidence scores while keeping the encoder and fusion rule fixed. The private banks allow each modality to express task-relevant information in a modality-specific subspace, and the shared anchors convert the adapted representation into a confidence score tied to the label structure. Anchor matching uses $\tilde{h}_i^m$ directly, without a residual connection to h_i^m. Thus, SEER-L1 modifies how modality confidence is estimated rather than adding a new backbone or fusion operator.

To further align the shared anchors with the downstream task, we add prototype supervision.

Let $a(y_i)$ denote the target anchor index for sample i. For classification, this is the class label and for regression, it is the nearest ordered anchor. We define the prototype-supervision loss as

$$\mathcal{L}_{\text{proto}} = \frac{1}{N|\mathcal{M}|} \sum_{i=1}^{N} \sum_{m \in \mathcal{M}} \text{CE}(\gamma_i^m, a(y_i)) \,,$$

where $\text{CE}(\cdot, \cdot)$ denotes the cross-entropy loss.

The prototype loss complements the contrastive term. The contrastive objective encourages label-consistent neighborhoods in the modality representation space, while prototype supervision aligns the shared-anchor assignments with the target labels.

3.4 Optional prototype-attentive evidence extraction

We also evaluate SEER-L2, which uses the shared anchors for temporal evidence extraction. Instead of relying on the pooled feature h_i^m, SEER-L2 allows the shared anchors to attend over the full modality sequence H_i^m and uses the resulting prototype-guided evidence vector in downstream computations. This extension tests whether replacing last-state pooling with anchor-guided temporal evidence improves the routing pipeline. Since SEER-L2 changes temporal evidence extraction together with routing and confidence estimation, we evaluate it separately from the main EAMC-SEER-L1 comparison.

3.5 Training objective and variant instantiation

Training combines the fused task loss, unimodal auxiliary supervision, inherited routing and distillation regularizers from the shared backbone implementation, and the SEER-specific losses described above. In compact form, the objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{proto}}\mathcal{L}_{\text{proto}} + \mathcal{L}_{\text{aux}},$$

where $\mathcal{L}_{\text{task}}$ denotes the fused prediction loss together with unimodal supervision, $\mathcal{L}_{\text{con}}$ is the label-aware contrastive loss used in SEER-L0, $\mathcal{L}_{\text{proto}}$ is the prototype-supervision loss used in SEER-L1, and $\mathcal{L}_{\text{aux}}$ contains the inherited routing and distillation regularizers. Variants that do not use a component set the corresponding coefficient to zero.

This objective gives a common training template for the model variants. EAMC uses encoded-space prototype routing without the SEER-specific losses. SEER-L0 adds the label-aware contrastive term. SEER-L1 adds shared-private prototype routing and prototype supervision. SEER-L2 is evaluated separately because it replaces pooled modality representations with prototype-guided temporal evidence. Optimization settings and dataset-specific hyperparameters are described in Section 4.

4 Experiments

We evaluate the proposed model variants on two multimodal sentiment benchmarks, CMU-MOSI and CMU-MOSEI, and one multimodal intent benchmark, MIntRec. The experiments are designed to answer three questions: whether estimating modality reliability in encoded feature space improves over raw-feature routing, whether label-aware supervision and label-structured anchor matching provide additional benefit, and whether the observed changes are better explained by confidence estimation rather than by added temporal or fusion complexity.

4.1 Datasets and reporting protocol

CMU-MOSI and CMU-MOSEI are used for sentiment regression with aligned features, while MIntRec is used for intent classification. Under the current preprocessing pipeline, CMU-MOSI contains 1,284 training samples, 229 validation samples, and 686 test samples; CMU-MOSEI contains 16,326 training samples, 1,871 validation samples, and 4,659 test samples; and MIntRec contains 1,334 training samples, 445 validation samples, and 445 test samples.

For MOSI and MOSEI, we report 7-class accuracy (Acc-7), binary accuracy (Acc-2), binary F1, and mean absolute error (MAE). For MIntRec, we report accuracy, Weighted F1, precision, and recall. To facilitate comparison with Fang et al. (2025), we follow the per-run peak-test convention used in EMOE: for MOSI and MOSEI, each run is summarized by the epoch with the best binary F1, and for MIntRec, by the epoch with the best Weighted F1. Our aggregation differs from EMOE in that we report mean ± standard deviation over three completed runs after applying the same per-run selection rule. This retains the comparison protocol while also reporting run-to-run variation. Unless otherwise noted, all local results use this three-run aggregation.

4.2 Implementation details

All local variants use the same multimodal backbone as the as the reproduced EMOE comparison baseline (Fang et al., 2025). Text is encoded with BERT Devlin et al. (2019), audio and video are projected to a shared hidden dimension before modality-specific Transformer encoding (Vaswani et al., 2017), and fusion uses the sum-style fused head. Unless otherwise stated, the shared hidden size is 256, each modality encoder uses 8 attention heads, the encoder depth is 4, and the batch size is 16. EAMC uses encoded-space prototype-conditioned routing with eight learnable prototypes and temperature 0.07. SEER-L0 keeps the same router and changes only the supervision. SEER-L1 replaces the prototype bank with shared-private anchor matching, and SEER-L2 adds prototype-guided temporal evidence extraction. The main comparison focuses on EAMC, SEER-L0, and SEER-L1, while SEER-L2 is reported as a separate extension.

Table 1: Main results on aligned CMU-MOSI. Published baseline results are shown above the separator and local three-run aggregates are shown below.

Method	Acc-7 ↑	Acc-2 ↑	F1 ↑	MAE ↓
MulT (Tsai et al., 2019)	35.1	80.0	80.1	0.936
MISA (Hazarika et al., 2020)	41.8	84.2	84.2	0.754
Self-MM (Yu et al., 2021)	45.3	84.9	84.9	0.738
MMIM (Han et al., 2021)	45.8	84.6	84.5	0.717
DMD (Li et al., 2023)	46.2	83.2	83.2	0.721
EMOE (Fang et al., 2025)	47.7	85.4	85.4	0.7100
EAMC (ours)	45.72 ± 0.30	85.88 ± 0.75	85.80 ± 0.74	**0.7277 ± 0.0047**
SEER-L1 (ours)	44.22 ± 2.24	**85.98 ± 0.61**	**85.96 ± 0.58**	0.7396 ± 0.0268

4.3 Main results

Tables 1-3 report the main benchmark comparisons. The upper block of each table lists representative published baselines, including EMOE, while the lower block reports our local three-run aggregates for EAMC and SEER-L1. Since the two blocks are obtained under different reporting conditions, the published rows are used as external reference points. The main interpretation is based on the local in-family comparison between EAMC and SEER-L1. On the aligned sentiment benchmarks, SEER-L1 gives modest improvements over EAMC on the primary F1 metric. F1 increases from 85.80 to 85.96 on aligned CMU-MOSI and from 85.33 to 85.63 on aligned CMU-MOSEI; SEER-L1 also obtains the higher local Acc-2 in both cases. On MIntRec, EAMC and SEER-L1 are close overall: SEER-L1 gives the highest Weighted F1 and precision, while EAMC gives the highest accuracy and recall. These results support a narrower conclusion: within the evaluated model family, replacing the encoded-space prototype router with shared-private anchor matching

provides the most consistent improvement on the primary F1-style metrics. The qualitative analysis below is used as a diagnostic view and is not treated as the main evidence for the claim.

Table 2: Main results on aligned CMU-MOSEI. Published baseline results are shown above the separator and local three-run aggregates are shown below.

Method	Acc-7 ↑	Acc-2 ↑	F1 ↑	MAE ↓
MulT (Tsai et al., 2019)	52.3	82.7	82.8	0.572
MISA (Hazarika et al., 2020)	52.3	85.3	85.1	0.543
Self-MM (Yu et al., 2021)	53.2	84.5	84.3	0.540
MMIM (Han et al., 2021)	50.1	83.6	83.5	0.580
DMD (Li et al., 2023)	52.4	84.8	84.7	0.546
EMOE (Fang et al., 2025)	54.1	85.3	85.3	0.5360
EAMC (ours)	**52.52 ± 0.55**	85.41 ± 0.26	85.33 ± 0.22	**0.5426 ± 0.0074**
SEER-L1 (ours)	52.44 ± 0.50	**85.66 ± 0.36**	**85.63 ± 0.34**	0.5428 ± 0.0108

Table 3: Main results on MIntRec. Published baseline results are shown above the separator and local three-run aggregates are shown below.

Method	Acc ↑	Weighted F1 ↑	Precision ↑	Recall ↑
MAG-BERT (Rahman et al., 2020)	70.34	68.19	68.31	69.36
MulT (Tsai et al., 2019)	72.58	69.36	70.73	69.47
MISA (Hazarika et al., 2020)	72.36	70.57	71.24	70.41
EMOE (Fang et al., 2025)	72.58	70.73	72.08	70.86
EAMC (ours)	**72.88 ± 0.72**	73.01 ± 0.69	74.11 ± 0.93	**72.88 ± 0.72**
SEER-L1 (ours)	72.73 ± 1.69	**73.03 ± 1.46**	**74.23 ± 1.03**	72.73 ± 1.69

4.4 Layer-wise ablation

Table 4 compares EAMC through SEER-L2 using the primary F1-style metric for each dataset. This ablation identifies which routing component contributes most to the observed improvement. EAMC provides a competitive encoded-space routing baseline. SEER-L0 improves the mean F1 on CMU-MOSEI but not on CMU-MOSI or MIntRec, suggesting that label-aware contrastive supervision alone is not sufficient. SEER-L1 obtains the highest mean primary metric on all three datasets, although the margin over EAMC is small on MIntRec. SEER-L2 performs below SEER-L1 across the three benchmarks and is notably weaker on CMU-MOSI and MIntRec. These results indicate that the main improvement in this set of experiments comes from shared-private anchor matching rather than from prototype-guided temporal evidence extraction.

Table 4: Layer-wise ablation over the routing variants. MOSI and MOSEI report binary F1, while MIntRec reports Weighted F1 under the current evaluation protocol.

Model	MOSI F1 ↑	MOSEI F1 ↑	MIntRec Weighted F1 ↑
EAMC	85.80 ± 0.74	85.33 ± 0.22	73.01 ± 0.69
SEER-L0	85.49 ± 0.41	85.52 ± 0.05	72.51 ± 0.85
SEER-L1	**85.96 ± 0.58**	**85.63 ± 0.34**	**73.03 ± 1.46**
SEER-L2	84.50 ± 0.75	85.43 ± 0.20	69.25 ± 1.20

Figure 2 visualizes the same comparison. EAMC remains a competitive encoded-space baseline, SEER-L0 shows mixed changes, and SEER-L1 obtains the highest mean primary metric in all three panels. SEER-L2 does not retain the gains observed for SEER-L1 and performs substantially worse on CMU-MOSI and MIntRec.

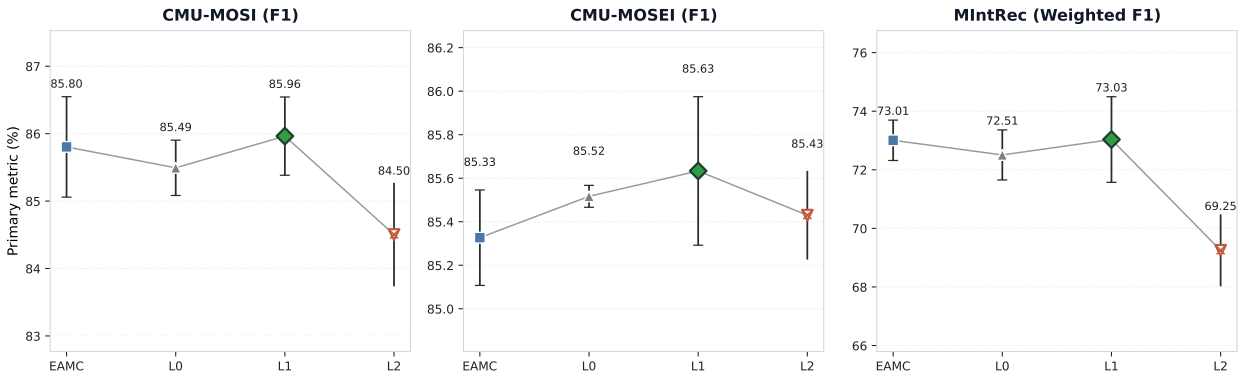

Figure 2: Layer-wise quantitative comparison on the three main datasets. Each point shows mean ± standard deviation over completed runs using binary F1 for aligned CMU-MOSI and aligned CMU-MOSEI, and Weighted F1 for MIntRec.

4.5 Qualitative representation progression

Figure 3 provides a qualitative view of the fused representations on the aligned CMU-MOSI test split. The figure uses a shared low-dimensional projection fitted on the same test examples across EAMC, SEER-L0, and SEER-L1 (van der Maaten & Hinton, 2008). Faint points denote individual utterances colored by sentiment value, and the overlaid centroid path summarizes the ordering of coarse sentiment bins from negative to positive.

This visualization is intended as an illustrative diagnostic rather than quantitative evidence. The centroid path appears more ordered for SEER-L1, while EAMC and SEER-L0 show less separation along the sentiment axis. This pattern is consistent with the quantitative results, but it should be interpreted only as qualitative support for the representation analysis.

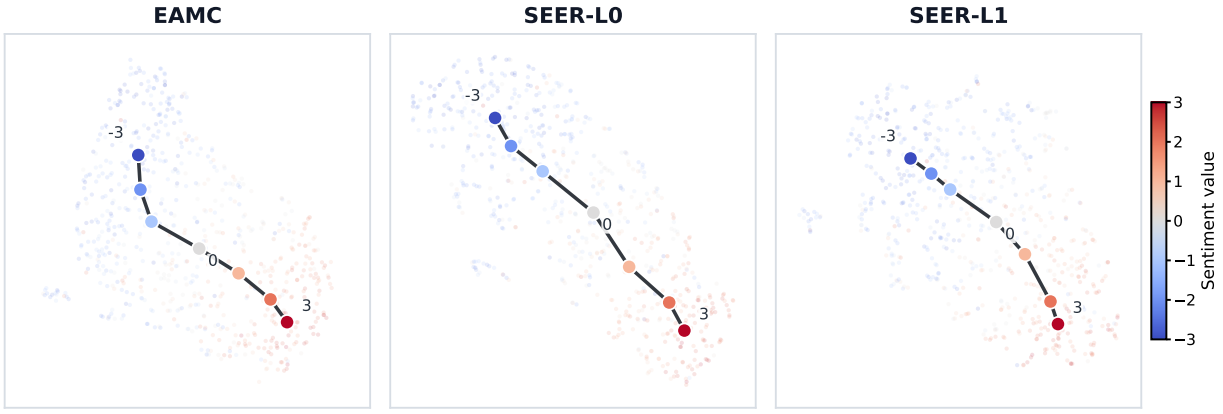

Figure 3: Qualitative comparison of fused representations on aligned CMU-MOSI. Faint points denote fused test representations for EAMC, SEER-L0, and SEER-L1 under a shared low-dimensional projection fitted on the same test examples across all three methods. The overlaid centroid path summarizes the ordering of coarse sentiment bins from negative to positive.

4.6 Parameter efficiency

The efficiency comparison is intended as an implementation-level observation rather than as a general architectural claim. In our reproduced EMOE implementation, the raw-feature router scales with the flattened

Table 5: Parameter counts by dataset, reported in millions.

	EMOE		EAMC		SEER-L1	
Dataset	Total	Non-BERT	Total	Non-BERT	Total	Non-BERT
MOSI	317.3	207.8	120.9	11.4	120.7	11.2
MOSEI	361.1	251.7	120.9	11.4	120.7	11.2
MIntRec	1123.8	1014.3	120.4	10.9	120.2	10.7

input dimensionality, whereas EAMC and SEER-L1 estimate modality reliability in encoded feature space. Table 5 reports the resulting total and non-BERT trainable parameter counts by dataset. The encoded-space variants use substantially fewer non-BERT routing and fusion parameters while remaining competitive on the primary metric.

On CMU-MOSI, reproduced EMOE uses 207.8M non-BERT trainable parameters, compared with 11.4M for EAMC and 11.2M for SEER-L1. On CMU-MOSEI, the corresponding counts are 251.7M, 11.4M, and 11.2M. Under our reproduced MIntRec feature configuration, the raw-feature EMOE router would scale to roughly 1.0B non-BERT trainable parameters, compared with 10.9M for EAMC and 10.7M for SEER-L1. We therefore interpret the MIntRec efficiency result as specific to this implementation and feature configuration, rather than as a published EMOE parameter count. A complementary performance-efficiency visualization is provided in Appendix B.

Additional results for unaligned settings and later-layer extensions are provided in Appendix A. We keep these results separate to maintain the main text's focus on the comparison from encoded-space routing to shared-private anchor-based confidence estimation.

5 Discussion

5.1 Interpreting the main results

The experiments indicate that the representation space used for routing has a measurable effect on adaptive multimodal fusion. EAMC shows that estimating modality reliability after semantic encoding is a competitive baseline. SEER-L0 adds label-aware contrastive supervision, but its effect is mixed across datasets. The main improvement appears in SEER-L1, where confidence estimation is changed from matching against an unconstrained prototype bank to matching modality-adapted representations against shared label-structured anchors.

This result suggests that adaptive fusion is most useful in this setting when the routing weights are estimated from task-structured evidence rather than from generic prototype confidence alone. Importantly, the improvement does not come from changing the encoder family or replacing the weighted-sum fusion rule. Instead, it comes from changing how modality confidence is computed before fusion. This supports the design choice of separating modality-specific adaptation from shared label-level comparison.

The evidence should be interpreted with some caution. On the sentiment benchmarks, the improvements are concentrated in Acc-2 and binary F1, while Acc-7 and MAE do not consistently favor SEER-L1. On MIntRec, EAMC and SEER-L1 are close: SEER-L1 gives the highest Weighted F1 and precision, whereas EAMC gives the highest accuracy and recall. We therefore view SEER-L1 as the best-supported variant on the primary F1-style metrics, not as a method that dominates across all reported measures.

5.2 Negative result and evaluation caveats

The SEER-L2 results clarify the role of temporal evidence extraction in the current setting. Prototype-attentive temporal evidence extraction is a plausible extension, since sentiment and intent cues may be localized within an utterance. However, SEER-L2 does not improve over SEER-L1 and performs substantially worse on CMU-MOSI and MIntRec. One possible explanation is that the modality encoders already

summarize sufficient temporal information for these benchmark configurations, so the additional temporal attention stage introduces complexity without improving the reliability estimate used for routing.

The comparison protocol also requires care. The closest adaptive-fusion baseline reports single-run peak test values (Fang et al., 2025). We follow the same per-run peak-test convention for comparability, but report mean and standard deviation over three completed runs. As a result, the local EAMC-SEER comparisons are more informative than direct headline comparisons with published single-run numbers. The parameter-efficiency comparison has a similar limitation: in our reproduced EMOE implementation, the raw-feature router scales with flattened input dimensionality. The resulting parameter gap should therefore be interpreted as specific to this implementation and feature configuration, rather than as a general property of all adaptive routing designs.

5.3 Limitations and future directions

This work has several limitations. First, the gains are moderate. They are obtained under a multi-run protocol and on top of a competitive adaptive-fusion baseline, but they should still be interpreted as incremental improvements rather than large empirical gains. Second, the experiments cover two English sentiment benchmarks and one English intent benchmark, with the main results focused on aligned feature settings. The conclusions may not transfer directly to multilingual data, strongly unaligned inputs, larger-scale pretraining setups, or tasks with substantially different modality statistics.

Third, the routing weights should not be interpreted causally. A high confidence value for a modality indicates stronger alignment with the learned label-structured routing space, not proof that the modality is uniquely responsible for the prediction. The interpretability claim is therefore limited to the representation and routing levels. Finally, the main paper focuses on EAMC, SEER-L0, and SEER-L1 because these variants most directly test the role of encoded-space routing and label-structured confidence estimation. Additional variants, including prototype-attentive temporal evidence extraction, relation-aware expert fusion, and robustness-oriented modality dropout, are reported in Appendix A; their results are more dependent on the dataset and metric.

Future work should examine the routing mechanism under broader conditions. Useful directions include evaluation on multilingual and additional intent or emotion datasets, robustness tests under partial modality corruption or asynchronous inputs, and alternative anchor structures for ordinal sentiment labels or uncertainty-aware intent categories.

6 Conclusion

This paper examined adaptive multimodal routing for sentiment analysis and intent recognition, focusing on how modality confidence is estimated before fusion. We introduced EAMC as an encoded-space routing baseline and SEER as a set of extensions that incorporate label structure into the routing representation and confidence-estimation mechanism. In particular, SEER-L0 adds label-aware contrastive supervision, while SEER-L1 replaces unconstrained prototype matching with shared-private anchor-based confidence estimation.

Experiments on aligned CMU-MOSI, aligned CMU-MOSEI, and MIntRec show that SEER-L1 provides the most consistent improvement over EAMC on the primary F1-style metrics, although the gains are moderate and not uniform across all reported measures. The ablation results also show that prototype-guided temporal evidence extraction does not improve performance in the current setting. These findings suggest that, for the evaluated benchmarks, improving the representation space used for modality confidence estimation is more useful than adding temporal pooling complexity.

References

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Kezhou Chen, Shuo Wang, Huixia Ben, Shengeng Tang, and Yanbin Hao. Mixture of multimodal adapters for sentiment analysis. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1822–1833, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.90. URL https://aclanthology.org/2025.naacl-long.90/.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, 2025.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022. URL http://jmlr.org/papers/v23/21-0998.html.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *ACM Multimedia 2024*, 2024. URL https://openreview.net/forum?id=78TMql1c04.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.723. URL https://aclanthology.org/2021.emnlp-main.723/.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 1122–1131, 2020.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 03 1991. ISSN 0899-7667. doi: 10.1162/neco.1991.3.1.79. URL https://doi.org/10.1162/neco.1991.3.1.79.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6631–6640, 2023.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, 2018.

Basil Mustafa, Carlos Riquelme Ruiz, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal contrastive learning with LIMoe: the language-image mixture of experts. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Qy1D9JyMBg0.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 2359–2369, 2020.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 6558–6569, 2019.

Laurens van der Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL https://api.semanticscholar.org/CorpusID:5855042.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 10790–10797, 2021.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pp. 1103–1114, 2017.

Hanlei Zhang, Hua Xu, Xin Wang, Qianrui Zhou, Shaojie Zhao, and Jiayan Teng. Mintrec: A new dataset for multimodal intent recognition. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 1688–1697, 2022.

Zhihong Zhu, Fan Zhang, Yunyan Zhang, Jinghan Sun, Zhiqi Huang, Qingqing Long, Bowen Xing, and Xian Wu. A survey on multi-modal intent recognition: Recent advances and new frontiers. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15223–15236, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.823. URL https://aclanthology.org/2025.findings-emnlp.823/.

A Additional Extensions and Supplementary Results

This appendix documents supplementary experiments and architectural extensions that broaden the design space explored in this work. The main text focuses on EAMC, SEER-L0, and SEER-L1 because this sequence directly tests whether label-structured modality confidence improves adaptive routing beyond encoded-space routing alone. The additional results reported here are useful for completeness, but they introduce extra modeling assumptions or produce more metric- and setting-dependent outcomes. We therefore treat them as supplementary evidence rather than as the basis of the main empirical result.

Unless otherwise noted, the appendix tables use the same per-run peak-selection and aggregation pipeline described in Section 4.1. The appendix documents the broader set of evaluated variants while keeping the main text focused on the EAMC-SEER-L1 routing comparison.

A.1 Additional extensions

Beyond the EAMC → SEER-L0 → SEER-L1 comparison, we investigated three further extensions. These variants are motivated by reasonable modeling hypotheses, but they address broader questions than the confidence-estimation mechanism emphasized in the main text. We therefore report them in the appendix rather than using them as the basis for the main comparison.

Prototype-attentive temporal evidence extraction (SEER-L2): SEER-L2 was motivated by the observation that emotional and intent-relevant evidence can be temporally sparse. If temporal evidence extraction is the limiting factor, then allowing shared anchors to attend over the full modality sequence should improve over last-state pooling. In practice, however, the results are weaker and less stable than those of SEER-L1. This suggests that, for the current datasets and feature pipelines, anchor-guided temporal evidence extraction does not improve the reliability estimate used for routing. We therefore treat SEER-L2 as a supplementary negative result rather than as a central source of improvement.

Relation-aware expert fusion (SEER-L3): SEER-L3 was motivated by the idea that multimodal affective data often contain structured disagreement. In settings such as sarcasm, masked affect, or partial cross-modal inconsistency, a single weighted-fusion rule may be too restrictive. SEER-L3 therefore introduces a relation-aware expert mixture that conditions fusion on pairwise agreement statistics together with routing confidences. This extension remains plausible and sometimes competitive, especially in supplementary settings, but it also introduces additional assumptions at the fusion stage. As a result, it does not provide as direct a test of the paper's central routing hypothesis as the simpler EAMC to SEER-L1 progression.

Robustness-oriented modality dropout (SEER-L4): SEER-L4 was motivated by robustness rather than by a new routing mechanism. By applying modality dropout during training, the model is encouraged to redistribute confidence when one input stream is missing or unreliable. This extension is useful for probing robustness-oriented behavior and remains competitive in some supplementary settings, but its effect is again more conditional and setting-dependent than the core SEER-L1 result. We therefore treat it as a natural continuation of the design space rather than as part of the paper's main validated contribution. In this appendix, we report SEER-L4 under the same standard benchmark evaluation as the other variants. A dedicated missing-modality robustness study is left to future work.

Taken together, these three extensions show that additional modeling complexity does not consistently improve the main evaluation metrics. They broaden the explored design space, but the simpler EAMC-SEER-L1 comparison remains the most direct evidence for the role of label-structured confidence estimation.

A.2 Supplementary architecture of SEER-L3

SEER-L3 builds on the routing framework by replacing the single weighted-fusion rule with a relation-aware expert mixture. As illustrated in Figure 4, the module first computes pairwise agreement statistics across modalities, then combines those statistics with routing confidences to predict expert weights over three branches: consensus, complement, and conflict. The final output is formed by a weighted average of the

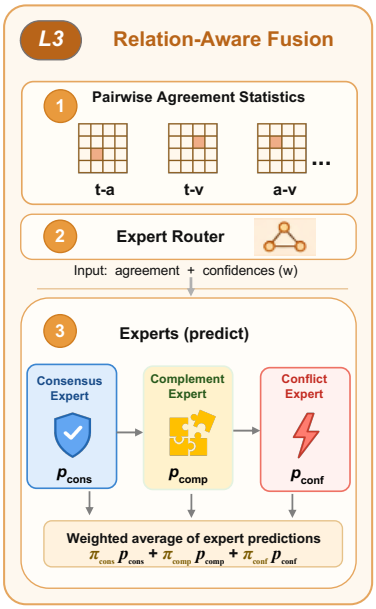

Figure 4: Supplementary architecture of SEER-L3. The relation-aware fusion module first computes pairwise agreement statistics across modalities, then uses agreement together with routing confidences to predict expert weights over three expert branches: consensus, complement, and conflict. The final output is formed by a weighted average of the expert predictions.

expert predictions. We include SEER-L3 in the appendix because it modifies the fusion stage in addition to the routing signal, making it less direct as a test of confidence-estimation design than the EAMC-SEER-L1 comparison.

Let μ_i and ν_i denote the mean and maximum pairwise agreement statistics for sample i, and let $\mathbf{c}_i$ denote the corresponding modality-confidence vector. The expert router computes

$$\boldsymbol{\pi}_i = \mathrm{softmax}(\mathrm{MLP}([\mu_i, \nu_i, \mathbf{c}_i])) , \tag{6}$$

and the final output is the weighted mixture

$$o_i = \pi_{i,\mathrm{cons}}\, o_i^{\mathrm{cons}} + \pi_{i,\mathrm{comp}}\, o_i^{\mathrm{comp}} + \pi_{i,\mathrm{conf}}\, o_i^{\mathrm{conf}} . \tag{7}$$

A.3 Aligned full-family comparison on the primary metric

Table 6 reports the aligned full-family comparison using the same primary metric emphasized in the main text: binary F1 for CMU-MOSI and CMU-MOSEI, and Weighted F1 for MIntRec. SEER-L1 obtains the highest mean primary metric across the three aligned benchmarks, while SEER-L3 and SEER-L4 remain competitive on the sentiment datasets. The later extensions do not consistently improve over SEER-L1, which is why they are treated as supplementary variants.

A.4 Supplementary results on unaligned CMU-MOSI

Table 7 reports the supplementary results on unaligned CMU-MOSI. Unlike the aligned setting, no single variant dominates every metric. EAMC gives the strongest Acc-2, SEER-L0 gives the strongest F1, SEER-L1 gives the strongest Acc-7, and SEER-L4 gives the best MAE. This setting therefore provides supplementary evidence of competitiveness, but it does not identify a single consistently best variant.

Table 6: Aligned full-family comparison on the primary metric. CMU-MOSI and CMU-MOSEI report binary F1; MIntRec reports Weighted F1. All rows are local aggregated results obtained under the same evaluation pipeline used in the main paper.

Model	MOSI F1 ↑	MOSEI F1 ↑	MIntRec Weighted F1 ↑
EAMC	85.80 ± 0.74	85.33 ± 0.22	73.01 ± 0.69
SEER-L0	85.49 ± 0.41	85.52 ± 0.05	72.51 ± 0.85
SEER-L1	**85.96 ± 0.58**	**85.63 ± 0.34**	**73.03 ± 1.46**
SEER-L2	84.50 ± 0.75	85.43 ± 0.20	69.25 ± 1.20
SEER-L3	85.62 ± 0.02	85.62 ± 0.25	72.24 ± 0.73
SEER-L4	85.61 ± 0.53	85.54 ± 0.10	72.06 ± 0.72

Table 7: Supplementary results on unaligned CMU-MOSI. Rows above the separator are representative paper baselines; rows below are local aggregated results under the same evaluation pipeline used in the main paper.

Method	Acc-7 ↑	Acc-2 ↑	F1 ↑	MAE ↓
EMOE (paper)	47.80	85.40	85.30	0.6970
EAMC (ours)	45.43 ± 1.09	**86.18 ± 0.92**	86.04 ± 0.93	0.7213 ± 0.0158
SEER-L0 (ours)	45.14 ± 0.42	86.13 ± 0.70	**86.07 ± 0.71**	0.7112 ± 0.0081
SEER-L1 (ours)	**46.94 ± 1.54**	85.57 ± 0.75	85.53 ± 0.73	0.7173 ± 0.0140
SEER-L2 (ours)	33.53 ± 15.69	85.06 ± 0.27	85.00 ± 0.33	0.9332 ± 0.2890
SEER-L3 (ours)	45.53 ± 0.37	85.57 ± 0.23	85.52 ± 0.19	0.7231 ± 0.0055
SEER-L4 (ours)	46.89 ± 1.11	85.93 ± 0.58	85.83 ± 0.58	0.6992 ± 0.0169

A.5 Supplementary results on unaligned CMU-MOSEI

Table 8 reports the supplementary results on unaligned CMU-MOSEI. In this setting, the later extensions remain competitive and the best metric values are distributed across multiple variants. In particular, SEER-L3 gives the strongest Acc-2 and F1 among the local variants, while EMOE remains the best external reference on Acc-7 and MAE. As with unaligned CMU-MOSI, these results document performance outside the aligned setting and should be interpreted as supplementary.

Table 8: Supplementary results on unaligned CMU-MOSEI. Rows above the separator are representative paper baselines; rows below are local aggregated results under the same evaluation pipeline used in the main paper.

Method	Acc-7 ↑	Acc-2 ↑	F1 ↑	MAE ↓
EMOE (paper)	53.90	85.50	85.50	0.5300
EAMC (ours)	52.51 ± 0.64	85.56 ± 0.17	85.47 ± 0.18	0.5430 ± 0.0107
SEER-L0 (ours)	53.15 ± 0.43	85.69 ± 0.25	85.65 ± 0.24	0.5380 ± 0.0032
SEER-L1 (ours)	52.90 ± 0.36	85.73 ± 0.02	85.69 ± 0.02	0.5395 ± 0.0050
SEER-L2 (ours)	50.72 ± 1.72	85.06 ± 0.40	84.77 ± 0.69	0.5605 ± 0.0208
SEER-L3 (ours)	53.27 ± 0.32	**85.99 ± 0.28**	**85.92 ± 0.28**	0.5367 ± 0.0060
SEER-L4 (ours)	52.48 ± 1.23	85.64 ± 0.22	85.55 ± 0.16	0.5432 ± 0.0107

A.6 Supplementary full-family results on MIntRec

Table 9 expands the main-paper MIntRec comparison to the full local method family. EAMC and SEER-L1 remain tightly clustered overall: SEER-L1 gives the strongest Weighted F1 and precision, while EAMC gives

the strongest accuracy and recall. SEER-L2 is substantially weaker than both. The later extensions remain competitive, but they do not improve over SEER-L1 on the primary Weighted F1 metric.

Table 9: Supplementary full-family results on MIntRec. Weighted F1 is reported explicitly under the current evaluation protocol. Rows above the separator are paper baselines; rows below are local aggregated results.

Method	Acc ↑	Weighted F1 ↑	Precision ↑	Recall ↑
MAG-BERT	70.34	68.19	68.31	69.36
MulT	72.58	69.36	70.73	69.47
MISA	72.36	70.57	71.24	70.41
EMOE (paper)	72.58	70.73	72.08	70.86
EAMC (ours)	**72.88 ± 0.72**	73.01 ± 0.69	74.11 ± 0.93	**72.88 ± 0.72**
SEER-L0 (ours)	72.73 ± 0.57	72.51 ± 0.85	73.25 ± 1.38	72.73 ± 0.57
SEER-L1 (ours)	72.73 ± 1.69	**73.03 ± 1.46**	**74.23 ± 1.03**	72.73 ± 1.69
SEER-L2 (ours)	69.51 ± 1.28	69.25 ± 1.20	70.95 ± 0.52	69.51 ± 1.28
SEER-L3 (ours)	72.51 ± 0.26	72.24 ± 0.73	72.90 ± 1.42	72.51 ± 0.26
SEER-L4 (ours)	72.06 ± 1.11	72.06 ± 0.72	73.03 ± 1.23	72.06 ± 1.11

A.7 Supplementary interpretation

The appendix results show that the broader SEER family remains competitive across additional settings, but the improvements are not uniform across datasets and metrics. SEER-L3 and SEER-L4 are useful extensions for exploring relation-aware fusion and robustness-oriented training, while SEER-L2 shows that prototype-guided temporal evidence extraction is not beneficial under the current configurations. These results support focusing the main text on EAMC, SEER-L0, and SEER-L1, where the comparison most directly isolates the effect of label-structured confidence estimation.

B Supplementary Parameter-Efficiency Visualization

The main text reports exact parameter counts in Table 5. Figure 5 provides a complementary view by plotting the primary F1-style metric against non-BERT trainable parameters. The visualization summarizes the performance-efficiency tradeoff, while the numerical comparison in Table 5 remains the primary reference.

As in Section 4.6, the MIntRec EMOE point should be interpreted as specific to our reproduced implementation and feature configuration rather than as a published EMOE parameter count.

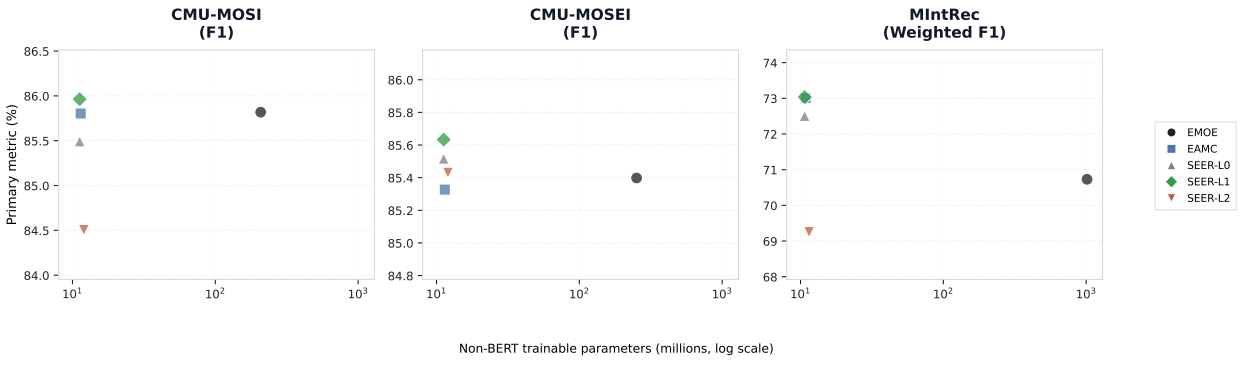

Figure 5: Supplementary performance-efficiency comparison. Each panel plots the primary F1-style metric against non-BERT trainable parameters on a log scale. The MIntRec EMOE point is configuration-specific, as described in Table 5.