FROM CUES TO SEGMENTS: LEARNING LATENT AFFECTIVE UNITS FOR MULTIMODAL TEMPORAL EMOTION LOCALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal temporal sentiment localization aims to identify and localize emotion-bearing segments in untrimmed multimodal sequences. However, cross-modal temporal misalignment and sparse affective cues are challenging for existing fusion methods, which often introduce redundant or unreliable information. To address these challenges, we propose LLAU, a multimodal fusion method that learns temporally structured latent affective units to construct dense and boundary-sensitive representations for emotion localization. Specifically, LLAU learns shared and modality-specific units to capture complementary affective cues. By jointly considering cross-modal support, semantic affinity and temporal affinity, it iteratively refines soft token-to-unit assignments and unit representations, thereby organizing sparse multimodal cues into compact affective representations. The learned assignments serve as affective priors for reconstructing dense features along the original timeline and guide modality-specific experts in weighting residual features to mitigate the loss of emotion boundary details during aggregation and reconstruction. We further introduce EmoTEL-8, a target-centric temporal emotion localization benchmark integrating visual, acoustic, and textual modalities. It provides identity-consistent visual tracks of target individuals, complete acoustic and textual context, and precise temporal boundaries for eight emotion categories. Experimental results demonstrate that LLAU achieves state-of-the-art performance on EmoTEL-8 and the public audio-visual benchmark TSL-300.

1 INTRODUCTION

Multimodal temporal emotion localization detects emotion-bearing segments in untrimmed videos, determines their temporal boundaries, and classifies their emotions, supporting applications such as human–robot interaction, risk screening, and adaptive education (Gordon et al., 2016; Majumder et al., 2019; Sawhney et al., 2020; Shukla et al., 2017).

In untrimmed videos, affective cues are sparse and much of the content is neutral. These cues may emerge asynchronously across modalities (Falagiarda & Collignon, 2019). Methods based on fixed temporal correspondence may therefore combine semantically unrelated affective cues, while modality-specific cues can provide complementary information (Hazarika et al., 2020; Fang et al., 2025). Effective fusion must therefore organize related evidence without relying solely on cross-modal support, while preserving modality-specific cues and local temporal variations needed for emotion boundaries.

Prior methods preserve modality-specific information, exploit cross-modal supervision, model segment boundaries, decompose shared and specific representations, or route dynamic experts (Zhang & Yang, 2022; Li et al., 2022; 2025; Hazarika et al., 2020; Fang et al., 2025). Cross-temporal interaction relaxes fixed temporal correspondence by allowing cues at different time steps to interact (Tsai et al., 2019). However, redundant or unreliable information may still dilute sparse affective cues. Our focus is on jointly modeling temporal associations and the roles of shared and modality-specific evidence within temporally structured latent affective units. We further use the learned assignments to guide dense reconstruction and the weighting of modality-specific residual features.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

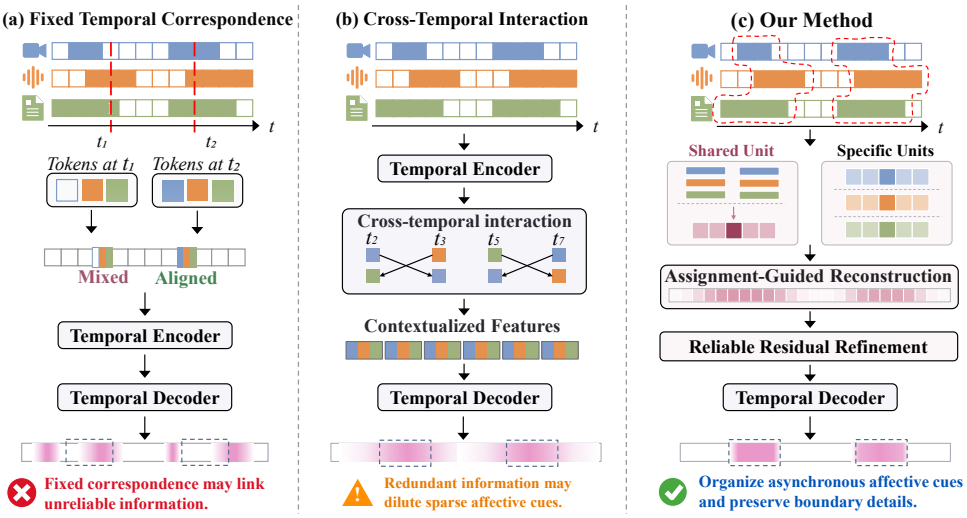

Figure 1: Comparison of LLAU with existing multimodal fusion strategies. (a) Fixed temporal correspondence may link unrelated affective cues. (b) Cross-temporal interaction may retain redundant or unreliable information. (c) LLAU combines temporally structured affective units with assignment-guided reconstruction and allocation-conditioned residual refinement.

We propose **LLAU**, a multimodal fusion method that learns temporally structured latent affective units. Shared and modality-specific units capture complementary cues, with temporal centers, window widths, and modality-dependent offsets encoding temporal structure. LLAU jointly uses cross-modal support, semantic affinity and temporal affinity to iteratively refine soft token-to-unit assignments and unit representations, accommodating local temporal misalignment while preserving modality-specific representations.

To preserve boundary details, LLAU derives affective priors from the final soft assignments for assignment-guided dense reconstruction along the original timeline. Modality-specific experts refine residuals between modality features and reconstructed representations. The same assignments condition residual gates and modulate expert contributions, allowing modality-specific local features to supplement the reconstructed representations. The resulting dense and boundary-sensitive representations can be integrated into existing temporal localization architectures. Figure 1 compares fusion under fixed temporal correspondence, cross-temporal interaction, and LLAU.

We also introduce **EmoTEL-8**, the first target-centric trimodal temporal emotion localization benchmark for untrimmed videos. It provides identity-consistent visual tracks as target-person inputs, complete acoustic and textual context, and uniformly annotated boundaries for eight emotion categories. At inference, models must localize and classify the designated person's intervals of all eight emotions without emotion queries or ground-truth boundaries.

LLAU achieves state-of-the-art performance on EmoTEL-8 and the public audio-visual benchmark TSL-300, with gains across localization architectures. Ablations examine the contributions of unit modeling, assignment-guided reconstruction, and modality-specific refinement.

Our main contributions are summarized as follows:

- We propose LLAU, which learns temporally structured latent affective units through iterative refinement of soft token-to-unit assignments and unit representations, jointly modeling temporal associations and the roles of shared and modality-specific evidence.
- We couple assignment-guided dense reconstruction with modality-specific refinement through final assignments that guide reconstruction and condition residual expert contributions.
- We introduce EmoTEL-8, a target-centric trimodal benchmark with identity-consistent visual tracks, complete acoustic and textual context, and precise boundaries for eight emotion categories.

Table 1: Comparison of temporal emotion localization datasets.

Dataset	Wild	V	A	T	Ann./video	Categories	Scale
MEMOS (Li et al., 2020)	×	✓	✓	–	1	5	40 sessions / 6,317 samples
TSL-300 (Zhang & Yang, 2022)	✓	✓	✓	–	4	2	300 videos / 1,642 samples
AVES (Li et al., 2025)	✓	✓	✓	–	7	6	167 videos / 1,747 samples
EmoTEL-8 (Ours)	✓	✓	✓	✓	**8**	**8**	**325 videos / 2,579 samples**

2 Related Work

2.1 Multimodal Fusion and Latent Representation Learning

In multimodal temporal emotion localization, many methods fuse audiovisual features at corresponding timestamps to localize emotions, including TSL-Net and BoCoNet (Zhang & Yang, 2022; Li et al., 2025). DCIN explores multiscale context and relaxed video–subtitle alignment, while FSENet enhances localization through face-guided interaction and boundary supervision (Li et al., 2022; Han et al., 2026).

General multimodal fusion methods employ various strategies to exploit complementary information. MulT models cross-temporal interactions, while MISA and EMOE emphasize shared/specific representations and dynamic experts, respectively (Tsai et al., 2019; Hazarika et al., 2020; Fang et al., 2025). These methods primarily target utterance- or clip-level affect prediction rather than emotion boundary localization. Compact representations offer another perspective: SA-MESH connects slot attention with optimal transport, while CLOT refines temporal features through frame–segment interaction and multistage transport (Zhang et al., 2023; Bueno-Benito & Dimiccoli, 2025). LLAU distinguishes itself through temporally structured shared and modality-specific affective units whose assignments jointly guide dense reconstruction and allocation-conditioned local refinement.

2.2 Emotion Datasets and Task Settings

Conventional benchmarks such as IEMOCAP, MELD, and DFEW primarily evaluate emotion recognition on pre-segmented utterances or clips, rather than interval discovery in untrimmed videos (Busso et al., 2008; Poria et al., 2019; Jiang et al., 2020). Person-level benchmarks specify whose emotion is analyzed under different task conditions. EMOTIC evaluates person-level recognition in images, MEmoR provides target emotion moments, and EmoTx predicts character-level emotions within movie scenes (Kosti et al., 2017; Shen et al., 2020; Srivastava et al., 2023). VEATIC evaluates continuous valence and arousal, whereas E^3 localizes person–emotion queries with the emotion category supplied at inference (Ren et al., 2024; Lin et al., 2024). Our setting instead requires localizing and classifying the designated person's emotion intervals without emotion queries or ground-truth boundaries. Temporal benchmarks progressively move from controlled emotion streams to in-the-wild interval localization. MEMOS provides participant-self-annotated emotion intervals from controlled multiplayer games (Li et al., 2020). TSL-300 localizes positive and negative sentiment in in-the-wild audio–visual videos, leaving the analyzed subject implicit in the video content (Zhang & Yang, 2022). AVES expands the label space to six basic emotions, with the subject specified by a fixed single-person view (Li et al., 2025). EmoTEL-8 instead provides identity-consistent visual tracks as explicit target-person inputs, retains complete acoustic and textual context, and annotates eight emotion categories with precise temporal boundaries. Table 1 compares dataset characteristics and annotation settings.

3 EmoTEL-8 Dataset

EmoTEL-8 is a target-centric benchmark for temporal emotion localization in untrimmed multimodal videos. Each video provides an identity-consistent visual track of a pre-designated person, together with global audio and textual context. The task is to localize and classify the target person's emotion intervals without predefined temporal boundaries or emotion queries. The benchmark covers eight target emotions: happiness, surprise, sadness, fear, disgust, anger, contempt, and confusion.

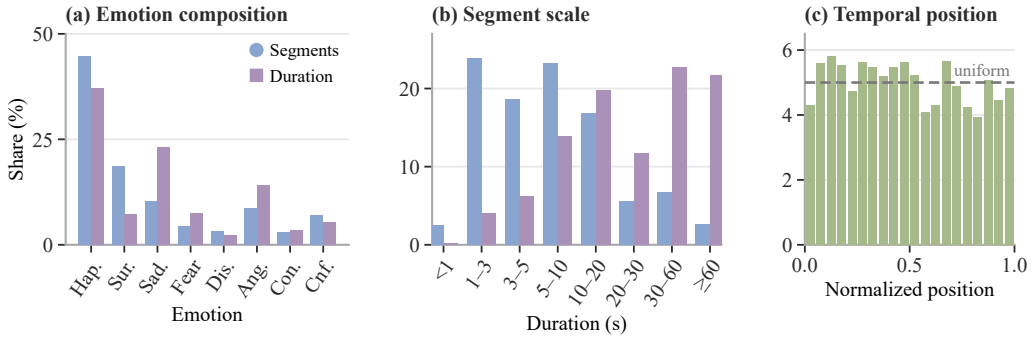

Figure 2: Statistics of EmoTEL-8: (a) class distributions by instance frequency and annotated duration; (b) segment-duration distributions; and (c) segment-center distributions along the normalized video timeline. All statistics are computed over the 2,579 intervals from the eight target emotion categories.

EmoTEL-8 contains 325 videos totaling approximately 22.3 hours, with 2,579 target-emotion intervals. We adopt a fixed video-level split of 211 training, 49 validation, and 65 test videos. Each video was independently annotated by eight annotators. Videos with Fleiss' $\kappa < 0.2$ were excluded. The mean per-video Fleiss' κ across the 325 retained videos was 0.47. Details of data collection, preprocessing, annotation, and split construction are provided in Appendix A.

Figure 2 summarizes the class composition, segment durations, and temporal positions of EmoTEL-8. Target-emotion intervals cover 38.16% of the total video duration, requiring models to identify emotional events within substantial background context. The dataset further exhibits a long-tailed class distribution and diverse temporal scales, encompassing both brief reactions and sustained emotional states. Emotion intervals occur throughout the videos rather than being concentrated within a narrow temporal region.

4 Methodology

For a sequence of length T, let $\mathcal{M} = \{\mathrm{v}, \mathrm{a}, \mathrm{t}\}$ denote the set of visual, acoustic, and textual modalities, with $M = |\mathcal{M}| = 3$. We omit the batch dimension. For each modality $m \in \mathcal{M}$, modality-specific LayerNorm and linear projections produce tokens $\mathbf{h}_m(t) \in \mathbb{R}^D$ at time steps $t = 0, \ldots, T-1$. The mask $M_{m,t} \in \{0, 1\}$ excludes missing observations and padding; invalid tokens are zeroed. Each modality accesses K shared units, K specific units, and one allocation sink, giving $J = 2K + 1$ destinations. Sink-assigned mass is excluded from the unit aggregation and reconstruction sums. The sink is not an affective unit; units need not match annotated intervals one-to-one.

LLAU refines units through adaptive allocation, then uses retained mass to guide dense reconstruction and residual refinement. A training-only emotion objective supervises the unit reconstruction (Figure 3).

4.1 Affective Unit Initialization

To accommodate local asynchrony, we measure cross-modal support within an integer temporal radius $\tau \geq 0$:

$$s_m(t) = \frac{1 + \max_{n \neq m,\, d \in \{-\tau, \ldots, \tau\}} \cos(\mathbf{h}_m(t), \mathbf{h}_n(t+d))}{2}. \tag{1}$$

Only valid, in-range candidates participate; invalid sources or empty candidate sets receive zero. Support measures cross-modal corroboration, rather than intrinsic validity of a modality. Let $n(t) = \sum_m M_{m,t}$ and $\mathbf{l}(t) = \sum_m M_{m,t}\mathbf{h}_m(t) / \max(1, n(t))$. The temporally pooled context of $\mathbf{l}$ predicts ordered centers c_k, widths w_k, and modality-dependent offsets $\delta_{m,k}$. Gaussian pooling initializes $\mathbf{u}_k^{\mathrm{sh},(0)}$ around c_k and $\mathbf{u}_{m,k}^{\mathrm{sp},(0)}$ around shifted centers. Geometry remains fixed during refinement.

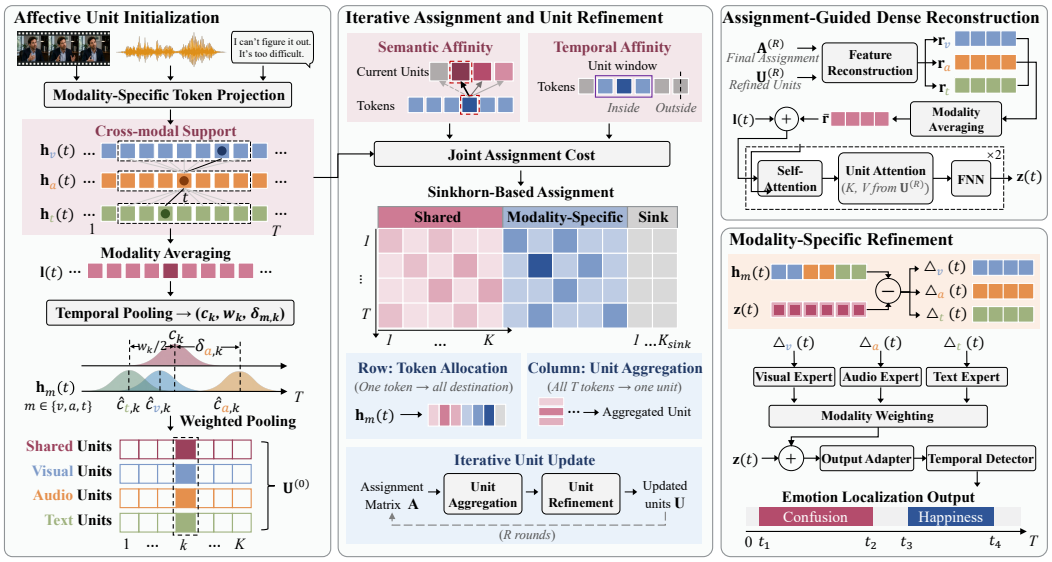

Figure 3: LLAU refines shared and specific units through adaptive allocation, then uses retained mass for dense reconstruction and residual refinement. A training-only emotion head supervises interval-pooled unit reconstructions.

4.2 Adaptive Assignment and Unit Refinement

At round $r = 0, \ldots, R-1$, semantic affinity, temporal proximity, and cross-modal support jointly determine the token–unit costs:

$$\begin{aligned} C^{\mathrm{sh},(r+1)}_{m,t,k} &= 1 - \cos\Big(\mathbf{h}_m(t), \mathbf{u}^{\mathrm{sh},(r)}_k\Big) + \lambda_{\mathrm{time}} C^{\mathrm{time}}_{m,t,k} + \lambda_{\mathrm{sh}}(1 - s_m(t)), \\ C^{\mathrm{sp},(r+1)}_{m,t,k} &= 1 - \cos\Big(\mathbf{h}_m(t), \mathbf{u}^{\mathrm{sp},(r)}_{m,k}\Big) + \lambda_{\mathrm{time}} C^{\mathrm{time}}_{m,t,k} + \lambda_{\mathrm{sp}} s_m(t). \end{aligned} \tag{2}$$

Here, $C^{\mathrm{time}}_{m,t,k}$ penalizes distance outside the unit window. Opposite support biases allow weakly corroborated but compatible tokens to enter specific units. Adding the learned sink cost $C^{\mathrm{sink}}_{m,t} = -\log p^{\mathrm{null}}_m(t)$ gives $\mathbf{C}^{(r+1)}_m \in \mathbb{R}^{T\times J}$.

For each nonempty modality, source masses are $\alpha_{m,t} = M_{m,t}/\sum_{t'} M_{m,t'}$. Suppressing the round index, let $\nu_{m,j} = \sum_t A_{m,t,j}$ be column masses and $\mu^g_m = \sum_{j\in\mathcal{J}_g} \nu_{m,j}$ their total in group $g \in \{\mathrm{sh}, \mathrm{sp}\}$, with $|\mathcal{J}_g| = K$. We fix source masses while allowing group totals to vary:

$$\begin{aligned} \min_{\mathbf{A}_m \geq 0,\, \mathbf{A}_m \mathbf{1} = \boldsymbol{\alpha}_m} \quad & \langle \mathbf{A}_m, \mathbf{C}_m \rangle + \varepsilon D_{\mathrm{KL}}\big(\mathbf{A}_m \,\big\|\, \boldsymbol{\alpha}_m (\mathbf{q}^{\mathrm{ref}})^\top\big) \\ & + \varepsilon \sum_{g\in\{\mathrm{sh},\mathrm{sp}\}} D_{\mathrm{KL}}\left(\boldsymbol{\nu}^g_m \,\middle\|\, \frac{\mu^g_m}{K}\mathbf{1}_K\right). \end{aligned} \tag{3}$$

Here, D_{KL} is generalized KL divergence, and $\boldsymbol{\nu}^g_m$ collects the group's column masses. The reference assigns $q^{\mathrm{ref}}_j = 1/(3K)$ to each non-sink destination and $q^{\mathrm{ref}}_{\mathrm{sink}} = 1/3$. This gives equal reference mass to the three groups without fixing their realized totals. The last term softly balances units within each non-sink group: its reference scales with the group's own mass, and the penalty vanishes for an unused group. No column marginal or sink quota is fixed. Differentiable generalized scaling and final row normalization approximate the objective; absent modalities have zero plans.

Non-sink assignments aggregate tokens into provisional units:

$$\begin{aligned} \widetilde{\mathbf{u}}^{\mathrm{sh},(r+1)}_k &= \frac{\sum_{m,t} A^{\mathrm{sh},(r+1)}_{m,t,k} \mathbf{h}_m(t)}{\sum_{m,t} A^{\mathrm{sh},(r+1)}_{m,t,k}}, \\ \widetilde{\mathbf{u}}^{\mathrm{sp},(r+1)}_{m,k} &= \frac{\sum_{t} A^{\mathrm{sp},(r+1)}_{m,t,k} \mathbf{h}_m(t)}{\sum_{t} A^{\mathrm{sp},(r+1)}_{m,t,k}}. \end{aligned} \tag{4}$$

Units with received mass at most ϵ_U use their initialization. Residual networks refine the provisional units, with shared units coupling modalities through subsequent costs. The final plans $\{\mathbf{A}_m^{(R)}\}_m$ and units $\mathbf{U}^{(R)}$ are used directly after R rounds.

4.3 Retention-Aware Dense Reconstruction

For valid tokens, define source-normalized allocations $\widetilde{A}^{g}_{m,t,k} = A^{g,(R)}_{m,t,k} / \max(\alpha_{m,t}, \epsilon_\alpha)$, $g \in \{\mathrm{sh}, \mathrm{sp}\}$, and set them to zero otherwise. The retained source-mass fraction is

$$p_m^{\mathrm{keep}}(t) = \mathrm{clip}\left(\sum_k \left(\widetilde{A}^{\mathrm{sh}}_{m,t,k} + \widetilde{A}^{\mathrm{sp}}_{m,t,k} \right), 0, 1 \right). \tag{5}$$

This fraction quantifies allocation to non-sink units, not an estimated probability that an observation is uncorrupted. Excluding the sink, we reconstruct each modality from the units:

$$\mathbf{r}_m(t) = \sum_k \widetilde{A}^{\mathrm{sh}}_{m,t,k} \mathbf{u}_k^{\mathrm{sh},(R)} + \sum_k \widetilde{A}^{\mathrm{sp}}_{m,t,k} \mathbf{u}_{m,k}^{\mathrm{sp},(R)}. \tag{6}$$

Let $\bar{\mathbf{r}}(t) = \sum_m M_{m,t} \mathbf{r}_m(t) / \max(1, n(t))$. The decoder query combines this unit reconstruction with a retention-weighted local timeline:

$$\mathbf{q}(t) = \frac{\sum_m p_m^{\mathrm{keep}}(t) \mathbf{h}_m(t)}{\max(1, n(t))} + \boldsymbol{\gamma} \odot \bar{\mathbf{r}}(t), \qquad \boldsymbol{\gamma} \in (0,1)^D. \tag{7}$$

Dividing by the observed-modality count preserves the effect of reduced retention.

The $K_U = (M+1)K$ units form memory $\mathbf{U}^{(R)} \in \mathbb{R}^{K_U \times D}$. Averaging shared allocations over valid modalities, concatenating the specific blocks, and normalizing over units gives prior $\boldsymbol{\Pi} \in \mathbb{R}^{T \times K_U}$. Decoder layers apply temporal self-attention, unit attention, and an FFN. Unit-attention head h computes

$$\mathbf{O}_h = \mathrm{MaskedSoftmax}_{\mathrm{units}}\left(\frac{\mathbf{Q}_h \mathbf{K}_h^\top}{\sqrt{d_{\mathrm{head}}}} + \lambda_h \log(\max(\boldsymbol{\Pi}, \epsilon_{\mathrm{attn}})) \right) \mathbf{V}_h, \tag{8}$$

where queries come from temporal self-attention, keys and values from unit memory, and $\lambda_h \in (0,1)$ is learned. Units receiving at most ϵ_U are masked. The projected multi-head update is scaled by $\bar{p}(t) = \sum_m p_m^{\mathrm{keep}}(t) / \max(1, n(t))$ before residual addition and LayerNorm; empty memory gives a zero update. The decoder outputs $\mathbf{z}(t) \in \mathbb{R}^D$.

4.4 Modality-Specific Residual Refinement

Experts recover local details from residuals $\boldsymbol{\Delta}_m(t) = \mathbf{h}_m(t) - \mathbf{z}(t)$. A content network predicts $b_m(t)$ from the reconstructed and modality features, their residuals and interactions, and retention and support. Normalized content weights are smoothed over valid modalities, then multiplied by retention:

$$\begin{aligned} a_m(t) &= \mathrm{MaskedSoftmax}_{m \in \mathcal{M}}(b_m(t); \{M_{m,t}\}_m), \\ g_m(t) &= p_m^{\mathrm{keep}}(t) \left[(1-\eta) a_m(t) + \eta \frac{M_{m,t}}{\max(1, n(t))} \right]. \end{aligned} \tag{9}$$

Here, $\eta = 0.05$ and invalid modalities receive zero weight. There is no subsequent renormalization: $\sum_m g_m(t) \le 1$, allowing all residual contributions to be attenuated when retention is low. We obtain

$$\mathbf{z}'(t) = \mathrm{LayerNorm}\left(\mathbf{z}(t) + \sum_m g_m(t) \lambda_m \mathcal{F}_m(\boldsymbol{\Delta}_m(t)) \right), \tag{10}$$

with modality-specific experts $\mathcal{F}_m$ and learned strengths $\lambda_m \in (0,1)$. A pointwise adapter outputs $\mathbf{Y} \in \mathbb{R}^{D_{\mathrm{out}} \times T}$ to the detector.

4.5 SEGMENT-SUPERVISED UNIT RECONSTRUCTION

During fully supervised training, an auxiliary classifier reads only the unit reconstruction $\bar{\mathbf{r}}(t)$, before the local query, decoder, or residual experts contribute. For interval $I_i = [s_i, e_i]$ in feature coordinates with label y_i, $\omega_i(t)$ measures overlap with feature bin $[t - \frac{1}{2}, t + \frac{1}{2}]$, masked by observation validity. We pool and train with

$$\mathbf{e}_i = \frac{\sum_t \omega_i(t)\bar{\mathbf{r}}(t)}{\sum_t \omega_i(t)}, \qquad \mathcal{L} = \mathcal{L}_{\text{det}} + \lambda_{\text{unit}}\mathcal{L}_{\text{unit}}, \qquad \lambda_{\text{unit}} = 0.1. \tag{11}$$

$\mathcal{L}_{\text{unit}}$ is the linear classifier's cross-entropy, averaged over usable intervals within each video and then over videos containing at least one usable interval; intervals without observed features are omitted. $\mathcal{L}_{\text{det}}$ is the original localization loss. The auxiliary branch supervises interval-level unit mixtures, without unit–interval matching or inferred background labels. Its GT inputs are confined to loss computation, and the classifier is discarded at inference. This loss is not used for video-level or point-level supervision.

Further methodological details are provided in Appendix B.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We use EmoTEL-8 as our primary benchmark, following the eight-category localization task and fixed data split described in Section 3. We use 1024-dimensional CycleACR (Chen et al., 2025) visual features, 1033-dimensional emotion2vec+ (Ma et al., 2024) audio features, and 1024-dimensional EmoBERTa (Kim & Vossen, 2021) textual features. LLAU projects each modality into a 256-dimensional space and produces 512-dimensional dense temporal features for downstream detectors. We report class-averaged mAP and instance-frequency-weighted wMAP at tIoU thresholds $\{0.1, 0.2, 0.3, 0.4, 0.5\}$, together with their averages across thresholds. We additionally evaluate LLAU on the audiovisual TSL-300 benchmark (Zhang & Yang, 2022), following its official 200/100 training–test split and using the officially released visual and acoustic features. Under video-level and point-level supervision, we report mAP at tIoU thresholds $\{0.10, 0.15, 0.20, 0.25, 0.30\}$, together with Average-mAP, Recall, and F_2 score, following the original evaluation protocol. Input features remain fixed during detector training on both datasets. Our experimental results are averaged over five random seeds; TSL-300 baseline results are taken from the cited papers. Metric definitions, feature extraction procedures, and additional implementation details are provided in Appendix C.

5.2 MAIN RESULTS ON EMOTEL-8

Comparison with multimodal fusion methods. Table 2 compares fusion methods within ActionFormer. Methods originally developed for other tasks are adapted to temporal emotion localization according to their core fusion designs. LLAU achieves the highest mean mAP and wMAP at all reported tIoU thresholds, with Average-mAP of 17.47 and Average-wMAP of 26.41, exceeding the strongest baseline, EMOE, by 1.48 and 2.93 percentage points, respectively. The gains persist at stricter tIoU thresholds, indicating stronger detection performance under tighter temporal-overlap requirements.

Evaluation across localization architectures. Table 3 compares LLAU with EMOE across ActionFormer (Zhang et al., 2022), TriDet (Shi et al., 2023), TemporalMaxer (Tang et al., 2023), and DyFADet (Yang et al., 2024a). LLAU improves Average-mAP by 1.48–2.48 percentage points and Average-wMAP by 1.98–2.93 percentage points over EMOE across the four detectors. These results support the applicability of LLAU across the four evaluated localization architectures.

5.3 EVALUATION ON THE AUDIOVISUAL TSL-300 BENCHMARK

To evaluate LLAU beyond the trimodal EmoTEL-8 setting, we train TSL-Net augmented with LLAU on the officially released audio-visual features of TSL-300. Table 4 compares audio-visual

Table 2: Comparison of multimodal fusion methods within ActionFormer on EmoTEL-8. The best and second-best results are shown in **bold** and underlined, respectively.

Method	mAP / wMAP at tIoU (%)					Average (%)	
	0.10	0.20	0.30	0.40	0.50	mAP	wMAP
GMU (Arevalo et al., 2017)	21.55 / 30.26	17.97 / 25.24	12.99 / 19.37	9.18 / 14.14	4.93 / 8.48	13.32 ± 1.18	19.50 ± 1.27
MulT (Tsai et al., 2019)	23.52 / 33.82	20.30 / 29.58	15.81 / 23.28	11.86 / 17.38	7.10 / 11.46	15.72 ± 0.60	23.10 ± 0.53
MTAG (Yang et al., 2021)	22.76 / 32.49	18.55 / 27.72	14.40 / 22.08	10.41 / 16.34	5.87 / 10.51	14.40 ± 1.05	21.83 ± 1.13
AcFormer (Zong et al., 2023)	22.53 / 33.00	18.89 / 27.84	14.53 / 22.04	10.79 / 16.57	5.62 / 10.43	14.47 ± 1.45	21.98 ± 1.82
DMD (Li et al., 2023)	21.69 / 31.43	18.41 / 26.84	14.01 / 21.38	9.91 / 15.18	5.73 / 9.79	13.95 ± 1.36	20.92 ± 1.50
GLFK (Yang et al., 2024b)	22.52 / 32.21	18.94 / 27.78	14.98 / 22.75	11.30 / 16.92	6.55 / 10.73	14.86 ± 1.02	22.08 ± 1.30
DLF (Wang et al., 2025)	22.26 / 33.36	19.48 / 29.45	15.48 / 23.28	11.58 / 17.46	6.52 / 11.76	15.06 ± 0.74	23.06 ± 0.50
DPDF-LQ (Zhou et al., 2025)	23.26 / 33.00	19.91 / 28.83	15.69 / 23.67	11.91 / 17.42	7.27 / 11.98	15.61 ± 1.40	22.98 ± 1.51
EMOE (Fang et al., 2025)	23.80 / 33.80	20.21 / 28.92	16.28 / 24.13	12.22 / 18.13	7.45 / 12.43	15.99 $\pm$ 0.85	23.48 $\pm$ 0.87
LLAU (Ours)	**25.07 / 36.35**	**21.73 / 32.51**	**17.90 / 27.70**	**13.58 / 20.81**	**9.08 / 14.68**	**17.47** $\pm$ 0.53	**26.41** $\pm$ 0.36

Table 3: Performance comparison of EMOE and LLAU with different downstream temporal detectors on EmoTEL-8.

Detector	Average-mAP (%)			Average-wMAP (%)		
	EMOE	LLAU	Δ (pp)	EMOE	LLAU	Δ (pp)
ActionFormer (Zhang et al., 2022)	15.99 ± 0.85	$\mathbf{17.47} \pm 0.53$	$+\mathbf{1.48}$	23.48 ± 0.87	$\mathbf{26.41} \pm 0.36$	$+\mathbf{2.93}$
TriDet (Shi et al., 2023)	13.48 ± 0.68	$\mathbf{15.96} \pm 0.54$	$+\mathbf{2.48}$	21.82 ± 1.30	$\mathbf{23.80} \pm 0.48$	$+\mathbf{1.98}$
TemporalMaxer (Tang et al., 2023)	14.85 ± 0.84	$\mathbf{16.69} \pm 1.12$	$+\mathbf{1.84}$	22.61 ± 0.86	$\mathbf{25.28} \pm 0.86$	$+\mathbf{2.67}$
DyFADet (Yang et al., 2024a)	10.68 ± 0.97	$\mathbf{12.73} \pm 1.23$	$+\mathbf{2.05}$	20.68 ± 1.31	$\mathbf{22.96} \pm 1.41$	$+\mathbf{2.28}$

methods under video-level and point-level supervision. Baseline results are taken from the audio-visual-only configurations reported by Zhang & Yang (2022) and Han et al. (2026).

The five-seed variation of our video-level and point-level results is reported in Appendix C.4. LLAU achieves the highest mAP at all five tIoU thresholds among the compared methods in both supervision settings. Under video-level supervision, it obtains 13.53 Average-mAP, outperforming UM by 2.78 percentage points, with an F_2 score of 30.47. Under point-level supervision, it achieves 22.17 Average-mAP, surpassing FSENet by 2.76 percentage points. These results demonstrate the effectiveness of LLAU in audio-visual localization under both weak-supervision regimes.

5.4 Ablation and Mechanism Analysis

Table 5 evaluates affective-unit modeling, adaptive assignment, auxiliary supervision, assignment-guided reconstruction, and modality-specific refinement on EmoTEL-8 with ActionFormer.

Affective-unit modeling. Replacing temporal structure with semantic queries removes unit geometry and temporal constraints, reducing Average-mAP and mAP@0.5 by 1.71 and 2.14 points. Shared-only units with a matched non-sink budget lower Average-mAP by 1.20 points; removing the sink lowers Average-mAP and mAP@0.5 by 0.83 and 1.12 points.

Parameter-matched control. The parameter-matched temporal network achieves 16.18 Average-mAP, 1.29 points below Full, suggesting that added capacity alone does not explain the gain.

Assignment and auxiliary supervision. Without unit-level auxiliary supervision, Average-mAP falls from 17.47 to 16.19. Under the same supervision, prior-aware row-softmax achieves 16.69. It preserves token–unit costs, destination priors, and source-row masses but removes within-group load regularization.

Residual refinement and assignment conditioning. Removing local residual refinement reduces Average-mAP and mAP@0.5 by 1.57 and 2.44 points. Restoring residual experts without assignment conditioning recovers 0.85 and 1.10 points; conditioning the residual branch adds another 0.72 and 1.34 points. With units, decoder attention, and experts retained, disabling assignment conditioning in both downstream branches yields 16.02 Average-mAP. Conditioning only reconstruction,

Table 4: Results on TSL-300 using audio-visual inputs. Bold marks the best reported result within each supervision setting; "–" indicates unreported metrics.

Supervision	Method	mAP@tIoU (%)					Average (%)		
		0.10	0.15	0.20	0.25	0.30	mAP	Recall	F_2 score
Video-level	CoLA (Zhang et al., 2021)	11.87	10.14	8.11	6.83	4.98	8.38	**64.07**	21.19
	UM (Lee et al., 2021)	14.77	13.28	10.77	8.41	6.54	10.75	55.33	23.94
	LLAU (Ours)	**21.20**	**16.43**	**12.51**	**9.94**	**7.56**	**13.53**	59.04	**30.47**
Point-level	LACP (Lee & Byun, 2021)	17.25	14.46	12.47	9.59	7.76	12.30	66.92	25.01
	SF-Net (Ma et al., 2020)	20.26	18.09	15.01	12.38	9.76	15.10	59.55	32.61
	HR-Pro stage 1 (Zhang et al., 2024)	16.64	13.95	11.86	9.74	8.01	12.04	–	–
	HR-Pro stage 2 (Zhang et al., 2024)	26.40	21.80	16.91	13.27	9.96	17.67	–	–
	TSL-Net (Zhang & Yang, 2022)	28.14	23.02	18.68	14.36	10.46	18.93	–	–
	FSENet (Han et al., 2026)	25.91	23.70	20.08	15.85	11.53	19.41	–	–
	LLAU (Ours)	**30.88**	**26.90**	**22.56**	**17.22**	**13.27**	**22.17**	**78.55**	**34.32**

Table 5: Ablation studies and controlled comparisons within ActionFormer on EmoTEL-8. All metrics are reported in percent.

Group	Variant	Average-mAP	Average-wMAP	mAP@0.5
Reference	**Full**	$\mathbf{17.47} \pm 0.53$	$\mathbf{26.41} \pm 0.36$	$\mathbf{9.08} \pm 0.49$
Unit modeling and matched alternatives	w/o Temporal Structure	15.76 ± 1.21	24.31 ± 1.13	6.94 ± 0.97
	Shared Units Only	16.27 ± 0.64	25.07 ± 0.53	7.39 ± 0.56
	w/o Sink	16.64 ± 0.48	24.74 ± 0.57	7.96 ± 0.31
	Parameter-matched Temporal Network	16.18 ± 0.58	24.62 ± 0.66	7.16 ± 0.51
Assignment and supervision	w/o Unit Auxiliary Supervision	16.19 ± 0.42	25.21 ± 0.58	7.80 ± 0.45
	Prior-aware Row-softmax	16.69 ± 0.47	25.64 ± 0.49	8.06 ± 0.42
Reconstruction and refinement	Standard Unit-Attention Decoder	16.50 ± 0.78	25.26 ± 0.98	7.21 ± 0.65
	w/o Local Residual	15.90 ± 0.50	24.86 ± 0.38	6.64 ± 0.52
	Residual w/o Assignment	16.75 ± 0.56	25.67 ± 0.63	7.74 ± 0.48
	Both Conditioning Paths Off	16.02 ± 0.64	24.78 ± 0.66	6.88 ± 0.52

only residual refinement, or both yields 16.75, 16.50, and 17.47, respectively. Average-wMAP and mAP@0.5 follow the same ordering.

These controls test distinct alternatives to the proposed mechanism. The parameter-matched network checks the effect of model capacity; Shared Units Only retains the non-sink unit budget while removing modality-specific units; and prior-aware row-softmax retains auxiliary supervision while removing within-group load regulation. Their lower detection scores suggest that the observed gain is not explained by any one of these factors in isolation.

Further implementation details and analyses appear in the appendix.

6 Conclusion

We presented LLAU, a multimodal fusion method that learns temporally structured shared and modality-specific latent affective units for temporal sentiment localization. By jointly considering cross-modal support, semantic affinity, and temporal affinity, LLAU iteratively refines soft token-to-unit assignments and unit representations. The final assignments provide affective priors for dense reconstruction along the original timeline and condition modality-specific residual contributions to the reconstructed features. We also introduced EmoTEL-8, a target-centric visual–acoustic–textual benchmark with identity-consistent target tracks and precise intervals for eight emotion categories. Experiments on EmoTEL-8 demonstrate consistent improvements across multiple temporal localization architectures. Further evaluation on the public audiovisual TSL-300 benchmark shows gains under both video-level and point-level supervision.

REFERENCES

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A González. Gated multimodal units for information fusion. *arXiv preprint arXiv:1702.01992*, 2017.

Elena Bueno-Benito and Mariella Dimiccoli. CLOT: Closed loop optimal transport for unsupervised action segmentation. In *ICCV*, pp. 10719–10729, 2025.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.*, 42(4):335–359, 2008.

Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, and Limin Wang. CycleACR: Cycle modeling of actor-context relations for video action detection. *IEEE TPAMI*, 47(11):10588–10603, 2025.

Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc. Natl. Acad. Sci. U.S.A.*, 114(38):E7900–E7909, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, volume 26, 2013.

Paul Ekman. An argument for basic emotions. *Cogn. Emot.*, 6(3-4):169–200, 1992.

Paul Ekman and Wallace V Friesen. A new pan-cultural facial expression of emotion. *Motiv. Emot.*, 10(2):159–168, 1986.

Federica Falagiarda and Olivier Collignon. Time-resolved discrimination of audio-visual emotion expressions. *Cortex*, 119:184–194, 2019.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. EMOE: Modality-specific enhanced dynamic emotion experts. In *CVPR*, pp. 14314–14324, 2025.

Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. Affective personalization of a social robot tutor for children's second language skills. In *AAAI*, volume 30, 2016.

Cailing Han, Zhangbin Li, Jinxing Zhou, Wei Qian, Jingjing Hu, Yanghao Zhou, Zhangling Duan, and Dan Guo. Face-guided sentiment boundary enhancement for weakly-supervised temporal sentiment localization. In *CVPR*, pp. 24759–24769, 2026.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM MM*, pp. 1122–1131, 2020.

Xingxun Jiang, Yuan Zong, Wenming Zheng, Chuangao Tang, Wanchuang Xia, Cheng Lu, and Jiateng Liu. DFEW: A large-scale database for recognizing dynamic facial expressions in the wild. In *ACM MM*, pp. 2881–2889, 2020.

Taewoon Kim and Piek Vossen. EmoBERTa: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*, 2021.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. Emotion recognition in context. In *CVPR*, pp. 1960–1968, 2017.

Pilhyeon Lee and Hyeran Byun. Learning action completeness from points for weakly-supervised temporal action localization. In *ICCV*, pp. 13628–13637, 2021.

Pilhyeon Lee, Jinglu Wang, Yan Lu, and Hyeran Byun. Weakly-supervised temporal action localization by uncertainty modeling. In *AAAI*, volume 35, pp. 1854–1862, 2021.

Juncheng Li, Junlin Xie, Linchao Zhu, Long Qian, Siliang Tang, Wenqiao Zhang, Haochen Shi, Shengyu Zhang, Longhui Wei, Qi Tian, and Yueting Zhuang. Dilated context integrated network with cross-modal consensus for temporal emotion localization in videos. In *ACM MM*, pp. 5083–5092, 2022.

Yan Li, Xiaohan Xia, Dongmei Jiang, Hichem Sahli, and Ramesh Jain. MEMOS: A multi-modal emotion stream database for temporal spontaneous emotional state detection. In *ICMI Companion*, pp. 370–378, 2020.

Yan Li, Wei Gan, Ke Lu, Dongmei Jiang, and Ramesh Jain. AVES: An audio-visual emotion stream dataset for temporal emotion detection. *IEEE TAC*, 16(1):438–450, 2025.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *CVPR*, pp. 6631–6640, 2023.

Wang Lin, Yueying Feng, Wenkang Han, Tao Jin, Zhou Zhao, Fei Wu, Chang Yao, and Jingyuan Chen. E^3: Exploring embodied emotion through a large-scale egocentric video dataset. In *NeurIPS*, volume 37, pp. 118182–118197, 2024.

Fan Ma, Linchao Zhu, Yi Yang, Shengxin Zha, Gourab Kundu, Matt Feiszli, and Zheng Shou. SF-Net: Single-frame supervision for temporal action localization. In *ECCV*, pp. 420–437, 2020.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of ACL*, pp. 15747–15760, 2024.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. DialogueRNN: An attentive rnn for emotion detection in conversations. In *AAAI*, volume 33, pp. 6818–6825, 2019.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *ACL*, pp. 527–536, 2019.

Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X Yu, and David Whitney. VEATIC: Video-based emotion and affect tracking in context dataset. In *WACV*, pp. 4455–4465, 2024.

Ramit Sawhney, Harshit Joshi, Saumya Gandhi, and Rajiv Shah. A time-aware transformer based model for suicide ideation detection on social media. In *EMNLP*, pp. 7685–7697, 2020.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. MEmoR: A dataset for multimodal emotion reasoning in videos. In *ACM MM*, pp. 493–502, 2020.

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. TriDet: Temporal action detection with relative boundary modeling. In *CVPR*, pp. 18857–18866, 2023.

Abhinav Shukla, Shruti Shriya Gullapuram, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Ramanathan Subramanian. Affect recognition in ads with application to computational advertising. In *ACM MM*, pp. 1148–1156, 2017.

Dhruv Srivastava, Aditya Kumar Singh, and Makarand Tapaswi. How you feelin'? learning emotions and mental states in movie scenes. In *CVPR*, pp. 2517–2528, 2023.

Tuan N. Tang, Kwonyoung Kim, and Kwanghoon Sohn. TemporalMaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023.

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 *ACL*, pp. 6558–6569, 2019.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. DLF: Disentangled-language-focused multimodal sentiment analysis. In *AAAI*, volume 39, pp. 21180–21188, 2025.

Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization. *IEEE TAC*, 9(2):255–270, 2018.

Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. MTAG: Modal-temporal attention graph for unaligned human multimodal language sequences. In *NAACL*, pp. 1009–1021, 2021.

Le Yang, Ziwei Zheng, Yizeng Han, Hao Cheng, Shiji Song, Gao Huang, and Fan Li. DyFADet: Dynamic feature aggregation for temporal action detection. In *ECCV*, pp. 305–322, 2024a.

Yang Yang, Xunde Dong, and Yupeng Qiang. CLGSI: A multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In *Findings of NAACL*, pp. 2099–2110, 2024b.

Can Zhang, Meng Cao, Dongming Yang, Jie Chen, and Yuexian Zou. CoLA: Weakly-supervised temporal action localization with snippet contrastive learning. In *CVPR*, pp. 16005–16014, 2021.

Chen-Lin Zhang, Jianxin Wu, and Yin Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, pp. 492–510, 2022.

Huaxin Zhang, Xiang Wang, Xiaohao Xu, Zhiwu Qing, Changxin Gao, and Nong Sang. HR-Pro: Point-supervised temporal action localization via hierarchical reliability propagation. In *AAAI*, volume 38, pp. 7115–7123, 2024.

Yan Zhang, David W Zhang, Simon Lacoste-Julien, Gertjan J Burghouts, and Cees GM Snoek. Unlocking slot attention by changing optimal transport costs. In *ICML*, pp. 41931–41951, 2023.

Zhicheng Zhang and Jufeng Yang. Temporal sentiment localization: Listen and look in untrimmed videos. In *ACM MM*, pp. 199–208, 2022.

Miao Zhou, Lina Yang, Thomas Wu, Dongnan Yang, and Xinru Zhang. Dual-path dynamic fusion with learnable query for multimodal sentiment analysis. In *EMNLP*, pp. 11355–11365, 2025.

Daoming Zong, Chaoyue Ding, Baoxiang Li, Jiakui Li, Ken Zheng, and Qunyan Zhou. ACFormer: An aligned and compact transformer for multimodal sentiment analysis. In *ACM MM*, pp. 833–842, 2023.

A Additional Details of EmoTEL-8

A.1 Data Collection and Preprocessing

We assembled an initial pool of 1,000 candidate videos from films and television programs, livestreams, and personal vlogs, including selected original videos from AVES (Li et al., 2025) and Ekman6 (Xu et al., 2018). The collection spans fixed-camera, single-person settings as well as dynamic scenes and multi-person interactions, providing diverse visual and conversational contexts for target-centric emotion localization. All retained videos were newly annotated with emotion categories and temporal boundaries under a unified target-centric protocol.Figure 4 illustrates this task setting through representative multimodal examples from EmoTEL-8.

We then screened candidate videos based on target-person observability, emotional variation, and contextual completeness. Retained videos had to provide sufficient cumulative screen time of the target person for annotators to identify emotion categories and temporal boundaries using audiovisual cues and surrounding context. Continuous on-screen presence was not required: videos containing camera motion, shot changes, flashbacks, or multi-person interactions were also eligible, provided that the observable portions of the target person supported these annotation requirements. Videos passing this initial screening were subsequently subjected to independent annotation and agreement-based quality control. After both content-based screening and annotation-agreement filtering, 325 untrimmed videos were retained to constitute EmoTEL-8. Details of the agreement-based filtering procedure are provided in the annotation protocol and quality control section.

To construct consistent trimodal inputs, we designated one target person per video before emotion annotation. For videos with camera motion or multiple people, we manually annotated identity-consistent person tracks over the intervals in which the target was visible; for fixed-camera, single-person videos, we used the full frame as the target region. All videos were standardized to 24 fps while retaining their original spatial resolution. Audio was loudness-normalized, resampled to 16 kHz, and converted to mono. The text stream consists of timestamped transcripts containing speaker labels, utterance start and end times, and utterance content. Non-English transcripts were automatically translated into English, with the original-language versions retained. While the visual input focuses on the designated person, the audio and text streams preserve the global interaction context to support interpretation of that person's emotions.

A.2 Annotation Protocol and Quality Control

EmoTEL-8 builds on Ekman's six basic emotions—happiness, surprise, sadness, fear, disgust, and anger—(Ekman, 1992) and adds contempt, supported by evidence of cross-cultural recognition (Ekman & Friesen, 1986), and confusion, which captures affective responses associated with uncertainty (Cowen & Keltner, 2017). Together, these form eight target emotion categories. Segments that clearly convey emotions outside this taxonomy receive an auxiliary *other* label, distinguishing non-target emotions from neutral states without expanding the target category set. These auxiliary labels are excluded from the training and evaluation annotations for the eight-class task. Annotators labeled only the pre-designated target person's emotions, recording the category, start time, and end time of each segment based on facial expressions, body movements, vocal behavior, and the surrounding audiovisual and conversational context.

Given the subjectivity of emotion perception, we recruited eight graduate students, including at least four with two or more years of research experience in affective computing. Before formal annotation, all annotators studied a common annotation guideline and completed pilot annotations. Disagreements arising during the pilot phase were discussed to establish a shared understanding of the emotion categories, target-person identification, and temporal-boundary criteria. Each of the eight annotators then watched every video in full to understand the target person and surrounding context before independently identifying emotion segments and assigning their categories and temporal boundaries. Relevant segments and their surrounding context were reviewed repeatedly when necessary. The resulting eight independent annotations per video were subsequently collected for agreement assessment and quality review.

Following the agreement-based video filtering procedure of AVES (Li et al., 2025), we computed Fleiss' κ separately for each video using the eight independent annotations. Videos with $\kappa < 0.2$ were excluded before joint discussion and final adjudication. After content screening and agreement-

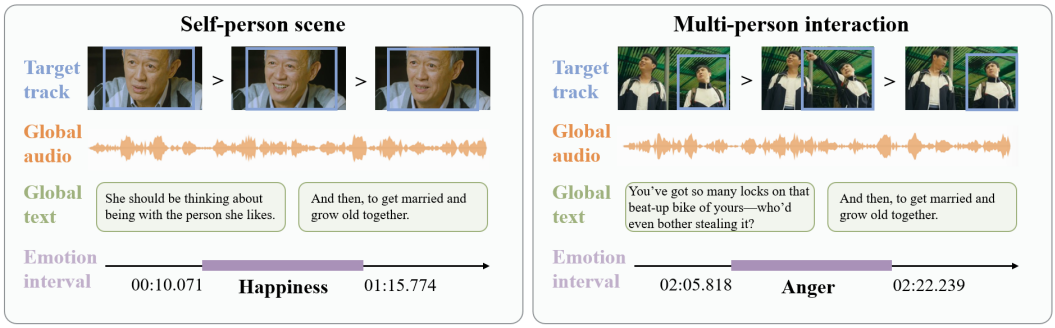

Figure 4: Task definition and representative examples of EmoTEL-8. Given an identity-consistent target track together with global audio and timestamped text, models localize and classify every emotion interval attributed to the designated person. English translations are shortened for readability.

Table 6: Split statistics of EmoTEL-8. Instance counts include only the eight target emotion categories used in the main evaluation.

Split	Videos	Duration (hh:mm:ss)	Target instances
Training	211	14:33:33	1,665
Validation	49	03:27:36	403
Test	65	04:17:41	511
Total	325	22:18:50	2,579

based filtering, 325 videos were retained. Across these retained videos, the unweighted arithmetic mean of the pre-adjudication Fleiss' κ coefficients was 0.47. Senior annotators then reviewed the retained annotations. Segments with remaining disagreements over emotion categories or temporal boundaries were jointly reviewed using the target person's multimodal cues and surrounding context, and their final categories and boundaries were determined through discussion.

A.3 Detailed Dataset Statistics and Split Construction

Unless otherwise specified, emotion statistics in this paper refer to the 2,579 intervals from the eight target emotion categories.[1] The distributions in Figure 2 follow this same convention. The target-emotion intervals span 30,651.991 seconds out of 80,330 seconds of video. Across all 325 videos, the mean number of target-emotion intervals is 7.94, with a median of 6 and a maximum of 45.

Figure 2(a) presents the class distribution in terms of both instance frequency and annotated duration. Happiness contains 1,151 instances, whereas Contempt contains 78. Sadness accounts for 10.31% of target-emotion instances but 23.10% of their cumulative duration. Instance shares are normalized by the 2,579 target-emotion intervals, and duration shares by their cumulative annotated duration.

Figure 2(b) shows the distribution of segment durations. Intervals shorter than 10 seconds account for 68.28% of all target instances. By contrast, intervals lasting at least 30 seconds constitute 9.27% of the instances but contribute approximately 44.3% of the annotated duration.

Figure 2(c) shows the distribution of segment centers along the normalized video timeline. For an interval with start and end times s_i and e_i in a video of duration D_v, its normalized position is $(s_i + e_i)/(2D_v)$. We divide the normalized timeline into 20 equal-width bins and report the percentage of target-emotion intervals whose centers fall within each bin.

[1]The complete annotation set additionally contains 79 intervals labeled *Other*, bringing the total to 2,658 affective intervals. These auxiliary intervals are excluded from the target-emotion statistics.

To reduce potential information leakage across splits, we group videos according to recorded associations in person identity and media source, keeping each group within a single partition. Subject to this constraint, the split construction considers both video counts and the distributions of the eight target emotions.

We use a training, validation, and test setup with an approximate target video-count ratio of 65:15:20. The resulting split statistics are reported in Table 6. The training set is used for parameter learning, the validation set for hyperparameter tuning and model selection, and the test set for final evaluation.

A.4 Data availability

We will release the emotion annotations, target-person tracks, official data splits, and evaluation scripts for EmoTEL-8 to support reproducible evaluation.

B Additional Methodological Details

B.1 Affective Unit Initialization

Cross-modal support uses radius $\tau = 4$ and valid, in-range candidates. Let $v(t) = \mathbf{1}\{\sum_m M_{m,t} > 0\}$ denote observed positions and P_t the original temporal mask excluding padding. Physical length is $L = 1 + \max\{t : P_t = 1\}$, or zero for an empty sequence. Define

$$\begin{aligned} \kappa_L &= \max(L-1, 1), \quad \rho_t = t/\kappa_L, \quad \widehat{c}_{m,k} = c_k + \delta_{m,k}/\kappa_L, \\ \mathbf{g}_{\mathrm{ctx}} &= \frac{\sum_t v(t)\mathbf{l}(t)}{\max(1, \sum_t v(t))}, \qquad \bar{c}_k = \frac{k - \frac{1}{2}}{K}. \end{aligned} \tag{12}$$

Observation gaps retain their physical indices; only observed positions contribute to the pooled context. Geometry is predicted once:

$$\begin{aligned} c_k &= \mathrm{clip}\left(\bar{c}_k + \frac{\delta_c}{K}\tanh f_{c,k}(\mathbf{g}_{\mathrm{ctx}}), 0, 1\right), \\ w_k &= w_{\min} + (w_{\max} - w_{\min})\sigma(f_{w,k}(\mathbf{g}_{\mathrm{ctx}})), \\ \delta_{m,k} &= \delta_{\max}\tanh(f_{\delta,m,k}(\mathbf{g}_{\mathrm{ctx}})). \end{aligned} \tag{13}$$

We use $\delta_c = 0.45$, $(w_{\min}, w_{\max}) = (0.02, 0.15)$, and $\delta_{\max} = 4$ feature steps. Gaussian initialization is

$$\begin{aligned} G^{\mathrm{sh}}_{t,k} &= v(t)\exp\left[-\tfrac{1}{2}((\rho_t - c_k)/w_k)^2\right], \\ G^{\mathrm{sp}}_{m,t,k} &= M_{m,t}\exp\left[-\tfrac{1}{2}((\rho_t - \widehat{c}_{m,k})/w_k)^2\right], \\ \mathbf{u}^{\mathrm{sh},(0)}_k &= \frac{\sum_t G^{\mathrm{sh}}_{t,k}\mathbf{l}(t)}{\max(\epsilon_U, \sum_t G^{\mathrm{sh}}_{t,k})}, \\ \mathbf{u}^{\mathrm{sp},(0)}_{m,k} &= \frac{\sum_t G^{\mathrm{sp}}_{m,t,k}\mathbf{h}_m(t)}{\max(\epsilon_U, \sum_t G^{\mathrm{sp}}_{m,t,k})}, \end{aligned} \tag{14}$$

with $\epsilon_U = 10^{-6}$; geometry remains fixed during refinement.

B.2 Adaptive Assignment and Unit Refinement

Both shared and specific costs use

$$C^{\mathrm{time}}_{m,t,k} = \frac{1}{4}\min\left(4, \frac{[|\rho_t - \widehat{c}_{m,k}| - w_k/2]_+}{\max(w_k, 10^{-3})}\right), \tag{15}$$

where $[x]_+ = \max(x, 0)$. Thus w_k is the Gaussian scale and the full zero-cost window width. We set $(\lambda_{\mathrm{time}}, \lambda_{\mathrm{sh}}, \lambda_{\mathrm{sp}}) = (0.25, 0.20, 0.10)$. The sink score $p^{\mathrm{null}}_m(t) = \sigma(f_{\mathrm{null}}(\mathbf{h}_m(t)))$ is floored at 10^{-6} before taking its negative logarithm.

For nonnegative masses, the generalized KL and group penalty are

$$D_{\mathrm{KL}}(\mathbf{x}\|\mathbf{y}) = \sum_i \left[x_i \log \frac{x_i}{y_i} - x_i + y_i \right], \qquad \Omega(\boldsymbol{\nu}_m^g) = \sum_{j \in \mathcal{J}_g} \nu_{m,j} \log \frac{K \nu_{m,j}}{\mu_m^g}, \qquad \mu_m^g > 0. \tag{16}$$

We use the continuous zero-mass convention, including $D_{\mathrm{KL}}(\mathbf{0}\|\mathbf{0}) = 0$, and infinite divergence for positive mass against a zero reference; $\Omega(\mathbf{0}) = 0$. Penalties are computed per modality, with group totals free to vary.

Each of $R = 2$ refinement rounds uses $N_{\mathrm{sc}} = 20$ differentiable generalized scaling steps in float32 log space, following the scaling structure of entropic transport (Cuturi, 2013). Suppressing modality and refinement indices, set $\mathsf{H}_{tj} = \log q_j^{\mathrm{ref}} - C_{tj}/\varepsilon$, $\varepsilon = 0.07$, and $\psi_j^{(0)} = 0$. For $a = 1, \ldots, N_{\mathrm{sc}}$, update valid rows and each non-sink group:

$$\begin{aligned} \phi_t^{(a)} &= \log \alpha_t - \mathrm{LSE}_j \left(\mathsf{H}_{tj} + \psi_j^{(a-1)} \right), \\ \zeta_j^{(a)} &= \mathrm{LSE}_{t:\alpha_t>0} \left(\mathsf{H}_{tj} + \phi_t^{(a)} \right), \\ \log \widehat{\mu}_g^{(a)} &= 2\, \mathrm{LSE}_{j \in \mathcal{J}_g} \left(\frac{\zeta_j^{(a)} - \log K}{2} \right), \\ \psi_j^{(a)} &= \frac{\log \widehat{\mu}_g^{(a)} - \log K - \zeta_j^{(a)}}{2}, \quad j \in \mathcal{J}_g, \\ \psi_{\mathrm{sink}}^{(a)} &= 0. \end{aligned} \tag{17}$$

Here LSE is log-sum-exp; the group coefficient equals ε, giving the exponent $1/2$. Final row projection is

$$A_{t,j} = \alpha_t \, \mathrm{softmax}_j \left(\mathsf{H}_{t,j} + \psi_j^{(N_{\mathrm{sc}})} \right). \tag{18}$$

Invalid-row log scales use a -10^4 sentinel before final masking; invalid rows and absent modalities return zero plans. Valid row masses equal α_t up to floating-point precision. The group update solves its proximal subproblem exactly, but finite iterations only approximate the full optimum; final group totals need not equal intermediate $\widehat{\mu}_g$. Source normalization uses $\epsilon_\alpha = 10^{-12}$, and retained/sink fractions are clipped to $[0, 1]$.

Units receiving mass at most ϵ_U use their initialization; the resulting provisional units are refined as $\mathbf{u} = \widetilde{\mathbf{u}} + \mathcal{R}_g(\widetilde{\mathbf{u}})$, with separate shared and specific networks. Final units and the plan that produced them are used without another assignment step.

B.3 Retention-Aware Reconstruction and Residual Refinement

The channel-wise sigmoid query gate $\boldsymbol{\gamma}$ initializes to 0.25. Memory order is shared, visual-specific, acoustic-specific, then textual-specific. The assignment prior is

$$\begin{aligned} \widehat{\boldsymbol{\Pi}}_{t,:} = \mathrm{Concat}\Big(& \frac{\sum_m M_{m,t} \widetilde{\mathbf{A}}^{\mathrm{sh}}_{m,t,:}}{\max(1, n(t))}, M_{\mathrm{v},t} \widetilde{\mathbf{A}}^{\mathrm{sp}}_{\mathrm{v},t,:}, \\ & M_{\mathrm{a},t} \widetilde{\mathbf{A}}^{\mathrm{sp}}_{\mathrm{a},t,:}, M_{\mathrm{t},t} \widetilde{\mathbf{A}}^{\mathrm{sp}}_{\mathrm{t},t,:} \Big), \\ \Pi_{t,\ell} = & \frac{\widehat{\Pi}_{t,\ell}}{\max(10^{-8}, \sum_{\ell'} \widehat{\Pi}_{t,\ell'})}. \end{aligned} \tag{19}$$

Zero-mass rows remain zero. Shared allocations are averaged, while specific blocks are concatenated; their relative prior mass therefore depends on the observed-modality count. Memory slots are active only if their received mass exceeds ϵ_U: sum over modalities and time for shared slots, and time for specific slots. Prior entries are floored at 10^{-6} before logging, then inactive-memory logits are masked.

Two decoder layers use eight heads each, with learned $\lambda_h = \sigma(\theta_h)$. The complete projected unit-attention increment is

$$\bar{p}(t)\left[\mathbf{W}_o \operatorname{Concat}_h(\mathbf{O}_h(t)) + \mathbf{b}_o\right], \tag{20}$$

including the output bias, before residual addition and LayerNorm. Temporal attention and FFNs retain their standard residual updates. Temporal keys use $v(t)$; empty temporal or memory key sets temporarily retain one key. Empty-memory updates and all unobserved layer outputs are zeroed.

The residual-gate input is

$$[\mathbf{z}(t), \mathbf{h}_m(t), |\mathbf{\Delta}_m(t)|, \mathbf{z}(t) \odot \mathbf{h}_m(t), p_m^{\text{keep}}(t), s_m(t)] \longmapsto b_m(t). \tag{21}$$

Each modality has a $(4D + 2) \to 128 \to 1$ gate MLP and a $D \to 128 \to D$ expert. Sigmoid injection strengths λ_m initialize to 0.1. Equation (9) applies content softmax, uniform smoothing ($\eta = 0.05$), then retention multiplication, with no explicit log-retention prior or subsequent normalization. This constrains injection weights, not final feature norms. A pointwise linear projection, LayerNorm, and Dropout produce the detector input; unobserved positions are zeroed, and the union-observation mask accompanies the features.

B.4 Segment-Supervised Unit Reconstruction

The auxiliary head reads only the pre-decoder unit reconstruction $\bar{\mathbf{r}}(t)$. Training crops supply intervals $[s_i, e_i]$ in feature coordinates, with overlap weights

$$\omega_i(t) = v(t)\left[\min(t + \tfrac{1}{2}, e_i) - \max(t - \tfrac{1}{2}, s_i)\right]_+. \tag{22}$$

Usable intervals satisfy $\sum_t \omega_i(t) > 0$ and are pooled as in Eq. (11), with denominator floor 10^{-12}. A linear $D \to 8$ classifier gives $\widehat{\mathbf{y}}_i = \operatorname{softmax}(\mathbf{W}_{\text{cls}}\mathbf{e}_i + \mathbf{b}_{\text{cls}})$. For usable intervals $\mathcal{I}_b$ in video b and $\mathcal{B}_+ = \{b : |\mathcal{I}_b| > 0\}$, the reduction is

$$\mathcal{L}_{\text{unit}} = \frac{1}{|\mathcal{B}_+|} \sum_{b \in \mathcal{B}_+} \frac{1}{|\mathcal{I}_b|} \sum_{i \in \mathcal{I}_b} -\log \widehat{y}_{i,y_i}. \tag{23}$$

The loss is zero when $\mathcal{B}_+$ is empty; excluded intervals are counted for auditing. We use weight 0.1 without background or sink labels. Gradients reach the reconstruction, allocations, and upstream parameters. The classifier has $2{,}056$ parameters for $D = 256$ and is unused at inference; interval-free supervision omits this auxiliary loss.

C Additional Experimental Details

C.1 Evaluation Metrics

For EmoTEL-8, let AP_c^θ denote the Average Precision of emotion category c at tIoU threshold θ, and let N_c be its number of ground-truth instances in the evaluation set. With $C_{\text{emo}} = 8$ target categories, the instance-frequency weight is

$$w_c = \frac{N_c}{\sum_{c'=1}^{C_{\text{emo}}} N_{c'}}. \tag{24}$$

The class-averaged mAP and instance-frequency-weighted wMAP are

$$\text{mAP}^\theta = \frac{1}{C_{\text{emo}}} \sum_{c=1}^{C_{\text{emo}}} \text{AP}_c^\theta, \qquad \text{wMAP}^\theta = \sum_{c=1}^{C_{\text{emo}}} w_c \, \text{AP}_c^\theta. \tag{25}$$

For $\Theta_{\text{E}} = \{0.1, 0.2, 0.3, 0.4, 0.5\}$, the cross-threshold averages are

$$\text{Average-mAP} = \frac{1}{|\Theta_{\text{E}}|} \sum_{\theta \in \Theta_{\text{E}}} \text{mAP}^\theta, \qquad \text{Average-wMAP} = \frac{1}{|\Theta_{\text{E}}|} \sum_{\theta \in \Theta_{\text{E}}} \text{wMAP}^\theta. \tag{26}$$

For TSL-300, we follow its original protocol and report mAP at tIoU thresholds from 0.10 to 0.30 in increments of 0.05, together with Recall and F_2 score. The latter emphasizes recall through

$$F_\beta = \frac{(1 + \beta^2)\,\text{Precision} \times \text{Recall}}{\beta^2\,\text{Precision} + \text{Recall}}, \tag{27}$$

where $\beta = 2$. We report the average of each metric across the five tIoU thresholds.

C.2 Feature Extraction for EmoTEL-8

The three streams are represented on a common 0.5-second grid. We extract 1024-dimensional visual features using the pretrained CycleACR model (Chen et al., 2025) released by its authors, with a SlowFast-R101 backbone. Visual inputs follow the pre-designated person's track in multi-person or dynamic scenes; fixed-camera single-person scenes use the full frame as the target region. Audio and text retain the global interaction context.

For audio, emotion2vec+ (Ma et al., 2024) produces a 1024-dimensional representation and nine emotion-prediction scores, concatenated into a 1033-dimensional feature. For text, non-English utterances are automatically translated into English. EmoBERTa (Kim & Vossen, 2021) encodes the current utterance together with up to four preceding utterances into a 1024-dimensional representation. Utterances are not filtered by speaker, and each representation is assigned to the grid using the current utterance's timestamps. Locations without valid observations are zero-filled and marked by modality-specific validity masks; sequence padding is invalid in every modality. All feature extractors remain frozen during detector training.

C.3 Training Configurations and Detector Integration

On fully supervised EmoTEL-8, we optimize the detector's localization loss plus $0.1\mathcal{L}_{\text{unit}}$, without other auxiliary losses or training-time synthetic corruption. We use $K = 16$, $D = 256$, $D_{\text{out}} = 512$, one sink per modality, $R = 2$ refinement rounds, and 20 scaling steps. Both regularization terms in Eq. (3) share $\varepsilon = 0.07$. No fixed shared, specific, or sink mass is prescribed. The entropy reference assigns equal mass to the three groups and uniform weights within each non-sink group; realized group masses depend on allocation. The geometric bounds are $\delta_c = 0.45$, $(w_{\min}, w_{\max}) = (0.02, 0.15)$, and $\delta_{\max} = 4$ feature steps (two seconds).

Across the evaluated localization architectures on EmoTEL-8, we train LLAU and the detector for 30 epochs using AdamW with a learning rate of 10^{-4}, weight decay of 0.05, and batch size of 1. Five warm-up epochs precede cosine learning-rate annealing. We use exponential moving average and clip the gradient norm to 1.

On TSL-300, LLAU receives visual and acoustic features from the original TSL-Net encoders and injects its outputs through independent modality-wise residual gates before the localization head. We jointly train the augmented model from scratch for 2,000 optimization steps with a learning rate of 10^{-5} and batch size of 16, initializing the gate logits to -4. The localization-head architecture and post-processing remain unchanged.

C.4 Run-to-Run Variation on TSL-300

Table 7 reports the mean and standard deviation of LLAU over seeds 40–44 for the aggregate TSL-300 metrics. The literature baselines in Table 4 retain their originally reported results.

Table 7: TSL-300 results over five seeds (mean $\pm$ standard deviation, in percent).

Supervision	Average-mAP	Recall	F_2 score
Video-level	13.53 ± 1.06	59.04 ± 1.75	30.47 ± 1.23
Point-level	22.17 ± 0.64	78.55 ± 1.57	34.32 ± 0.85

C.5 Implementation of Ablations and Controlled Comparisons

This section provides additional implementation details for the ablation studies and controlled comparisons reported in Table 5, focusing on the modifications relative to Full LLAU.

Without unit auxiliary supervision. We set $\lambda_{\text{unit}} = 0$ while retaining the Full model's adaptive assignment, detector loss, architecture, training schedule, and checkpoint-selection protocol. The auxiliary classifier contributes no training gradients and is not used at inference in either configuration.

Table 8: Fusion-module parameter counts. Downstream detectors and frozen feature extractors are excluded.

Model	Parameters (M)
Full LLAU	3.921
Parameter-matched Temporal Network	3.929
Prior-aware Row-softmax	3.921

Prior-aware row-softmax. We replace the assignment optimization in Eq. (3) with

$$A_{m,t,j} = \alpha_{m,t}\,\mathrm{softmax}_j\left(\log q_j^{\mathrm{ref}} - C_{m,t,j}/\varepsilon\right). \tag{28}$$

Token–unit costs, destination priors, temperature, and source-row masses match Full. This change removes the within-group load regularization while leaving realized group masses free to vary. Both variants retain the same unit auxiliary loss with $\lambda_{\mathrm{unit}} = 0.1$.

Without explicit unit temporal structure. We replace temporally initialized shared and modality-specific units with learned semantic queries of the same counts. The predicted centers, widths, and modality-dependent offsets are removed, together with the temporal term in the token–unit cost. Semantic affinity, cross-modal support, destination types, unit refinement, and downstream modules remain as in Full.

Shared units only. We remove the modality-specific unit destinations and reallocate their unit budget to the shared group, matching Full's total non-sink unit budget. Each modality retains its allocation sink. Assignment, unit refinement, dense reconstruction, and residual refinement otherwise follow Full.

Parameter-matched temporal network. We replace unit initialization, assignment and refinement, assignment-based reconstruction, and unit attention with two temporal self-attention–FFN layers operating on the valid-modality mean. The FFN width is selected to match Full's fusion parameter count. Input projections, modality-specific residual experts, content gates, cross-modal support, and the output projection are retained. Without assignments, the retention inputs to residual gating are fixed to the non-sink prior mass.

Standard unit-attention decoder. We retain the unit memory, temporal self-attention, content-based unit attention, and residual experts. Decoder queries use the local timeline alone, the assignment-derived additive attention prior is removed, and the retained-mass multiplier $\bar{p}(t)$ on the unit-attention update is replaced by one. The residual branch retains assignment-conditioned gating and expert-output weighting.

Both conditioning paths off. We remove explicit final-assignment conditioning from both reconstruction and residual refinement. The decoder query uses the local timeline alone, and the additive attention prior is removed. The retained-mass multiplier on the unit-attention update is also replaced by one. Assignment-derived retention is replaced throughout residual gating by the fixed non-sink prior mass. Unit refinement, unit memory, content-based unit attention, and residual experts are retained; thus, unit content still depends on assignment. The single-branch controls apply the corresponding reconstruction or residual change alone. Parameters exclusive to the disabled reconstruction conditioning paths become inactive.

Parameter matching. Table 8 reports fusion-module parameter counts. The temporal control differs from Full by approximately 0.21%, while row-softmax introduces no additional learnable parameters.

Without the allocation sink. We remove the sink destination in every refinement round while keeping the shared and modality-specific unit counts unchanged. The reference weights are renormalized over the remaining destinations. All valid source mass is assigned to these units, giving $p_m^{\mathrm{keep}}(t) = 1$ for valid tokens. Training seeds and validation-based checkpoint selection match Full. This comparison evaluates null allocation together with its downstream retention-dependent effects.

Table 9: Fusion parameters and dense-operation counts.

Method	Fusion params. (M)	Dense GFLOPs
MulT	2.51	4.16
EMOE	1.91	1.95
LLAU	3.92	4.24

C.6 Computational Complexity

Table 9 compares fusion parameter counts and dense-operation costs for the evaluated implementations. LLAU has a dense-operation count close to MulT (4.24 versus 4.16 GFLOPs), but uses more parameters. Both its parameter count and dense-operation cost exceed those of EMOE, indicating an accuracy–cost trade-off rather than uniformly lower computational overhead.

D Additional Robustness and Mechanism Analysis

D.1 Qualitative Unit Mechanism

To make the role of LLAU units directly inspectable, we align representative video frames, ground-truth (GT) intervals, input availability, unit pooling, and decoder readout on a common video timeline. Red boxes indicate the tracked target person. In the input-availability rows, colored segments denote observed visual, acoustic, and textual inputs, whereas gray segments denote unavailable inputs. The pooling curve reports where a unit accumulates assignment mass at each 0.5-s bin, normalized by the unit's received mass over the whole video. The readout curve reports the decoder's share assigned to the selected unit at each 0.5-s bin, averaged over two decoder layers and eight attention heads and normalized across all 64 units. The bottom bar reports the GT-overlap-weighted mean decoder readout aggregated over the shared, visual, acoustic, and textual unit groups.

The examples illustrate four complementary behaviors. In Fig. 5(a), the shared unit S10 and the visual unit V10 are jointly read during a happiness event. Their whole-video pooling mass inside the target GT interval is 44.57% and 79.74%, respectively. In Fig. 5(b), visual unit V4 concentrates 92.70% of its whole-video pooling mass inside the target happiness interval.

In Fig. 6(c), text unit T16 aggregates evidence from separated utterances during an anger event. The unavailable-text interval remains visible and is not interpolated. T16 places 97.89% of its whole-video pooling mass inside the complete target GT interval. In Fig. 6(d), audio unit A10 receives evidence at multiple points throughout the complete confusion interval. Its whole-video pooling mass inside the GT is 74.62%, whereas its mean decoder readout within the GT interval is 7.05%. The varying readout curve indicates that the decoder consults A10 with time-varying strength across the event.

Taken together, these cases illustrate the intended mechanism of LLAU. Soft token-to-unit assignments organize sparse and temporally asynchronous observations into shared and modality-specific affective units. Shared units support joint readout when evidence is corroborated across modalities, whereas modality-specific units preserve reliable cues that are available only in one stream. Pooling indicates where a unit accumulates evidence, while decoder readout indicates when the downstream decoder consults that evidence. By retaining the original timeline and the input-availability masks, LLAU can use these structured units for assignment-guided dense reconstruction and modality-specific residual refinement without treating missing-input spans as observed evidence.

D.2 Contribution of Input Modalities

We evaluate LLAU with visual-only, visual–acoustic, visual–textual, and trimodal inputs on EmoTEL-8. Table 10 reports the results.

Adding either acoustic or textual context improves both metrics over visual-only inputs. Combining all three modalities further increases Average-mAP by 2.44 and 2.67 percentage points over visual–acoustic and visual–textual inputs, respectively. These results support the benefit of using both context modalities in the evaluated target-centric setting.

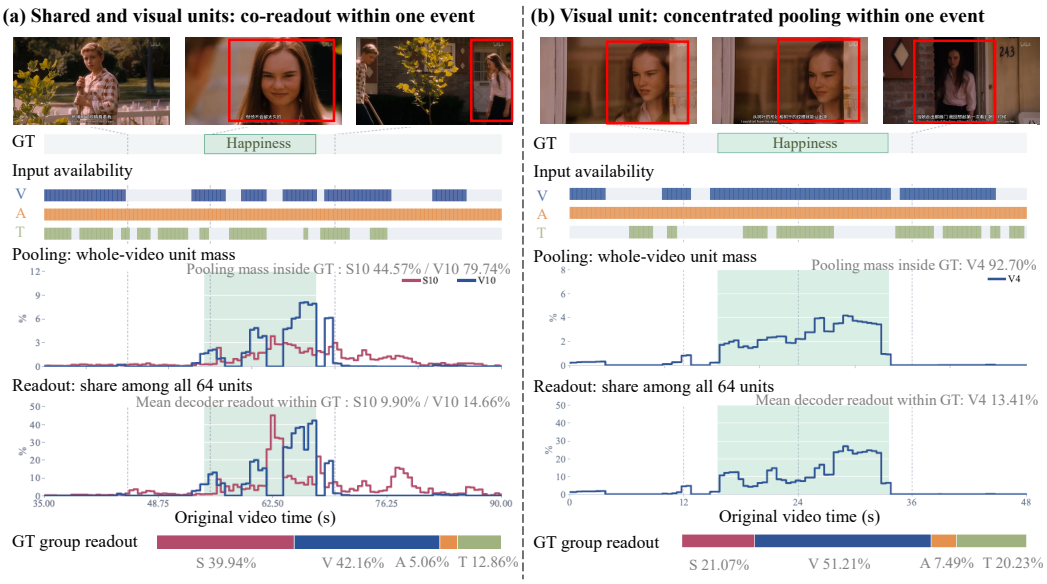

Figure 5: **Shared and visual unit behavior.** (a) Shared and visual units are jointly read during a happiness event. (b) A visual unit concentrates its whole-video pooling mass inside the target GT interval. Green shading marks the target GT, colored bars indicate input availability, and the bottom bars show GT-overlap-weighted mean group readout. Pooling and readout use separate normalizations.

Table 10: Input modality combinations on EmoTEL-8. V, A, and T denote visual, acoustic, and textual inputs.

Inputs	Average-mAP (%)	Average-wMAP (%)
V	12.14 ± 0.97	20.64 ± 1.23
V+A	15.03 ± 0.34	22.38 ± 0.74
V+T	14.80 ± 0.57	21.17 ± 0.80
V+A+T	$\mathbf{17.47} \pm 0.53$	$\mathbf{26.41} \pm 0.36$

D.3 Emotion Classification on Ground-Truth Intervals

To examine emotion-category information independently of interval discovery, we evaluate classification on the same ground-truth intervals using a common readout protocol. We compare LLAU with prior-aware row-softmax under the same auxiliary supervision, and with EMOE equipped with matched auxiliary classification supervision. Table 11 reports macro-F1 over the eight target emotion categories.

LLAU exceeds prior-aware row-softmax and EMOE by 4.51 and 2.81 percentage points, respectively. These results complement the detection metrics by assessing emotion discrimination within given intervals.

D.4 Controlled Temporal Misalignment

All results in this section are averaged over five random seeds.

We independently shift the acoustic or textual features and their validity masks by ± 0.5, ± 1, ± 2, and ± 4 seconds, while keeping visual features, emotion annotations, and sequence lengths unchanged. Out-of-range features are discarded, and newly vacated positions are zero-filled and marked invalid, without circular wraparound. Each magnitude includes four interventions: positive and negative shifts of either the acoustic or textual stream, with only one modality shifted at a time. For each

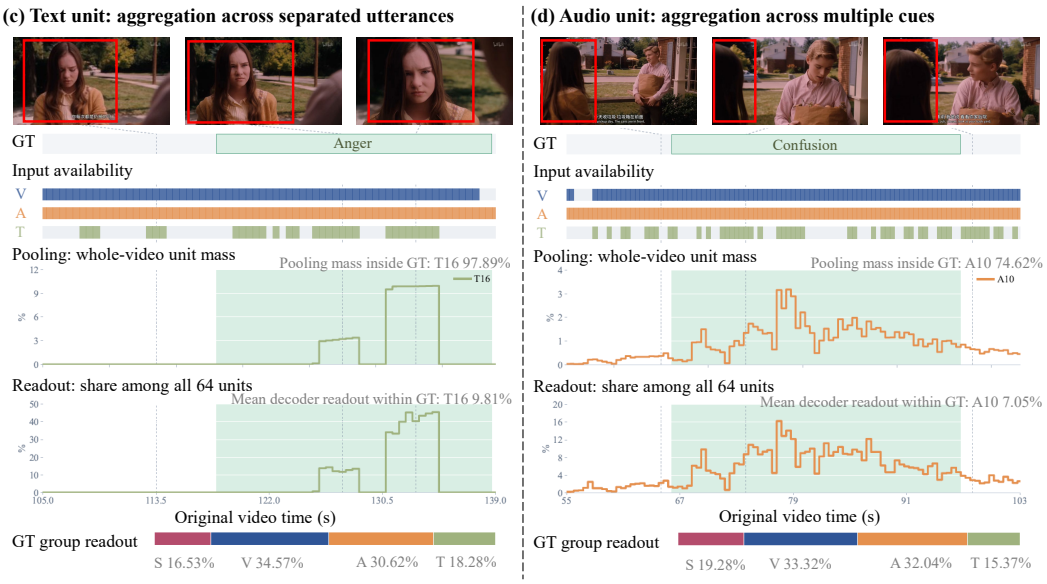

Figure 6: **Text and audio unit behavior.** (c) A text unit aggregates evidence from separated utterances while preserving the unavailable-text span. (d) An audio unit receives evidence at multiple points across the complete confusion GT and is consulted with time-varying strength. The curves show normalized unit pooling and decoder readout; missing-input intervals are shown without interpolation.

Table 11: Emotion classification on ground-truth intervals with matched auxiliary supervision. Values are mean ± standard deviation over five seeds; bold indicates the highest mean Macro-F1.

Method	Macro-F1 (%)
Prior-aware Row-softmax	31.10 ± 0.93
EMOE	32.80 ± 1.12
LLAU	$\mathbf{35.61} \pm 0.75$

seed, we average the Average-mAP drops relative to that model's clean-input performance across these four interventions, then report the mean over five seeds.

We compare full LLAU, *w/o Offset*, which removes only the learnable modality-dependent offsets, and the adapted GLFK baseline, which fuses modality features under fixed temporal correspondence (Yang et al., 2024b).

As shown in Table 12, all models exhibit small mean performance changes at ± 0.5 seconds. At magnitudes from ± 1 to ± 4 seconds, LLAU consistently shows the smallest mean degradation. At ± 4 seconds, its drop is 1.19 and 0.63 percentage points smaller than those of *w/o Offset* and GLFK, respectively. The former comparison supports the contribution of learnable modality-dependent offsets to misalignment robustness; the latter shows an advantage over the evaluated fusion baseline based on fixed temporal correspondence under this perturbation setting.

D.5 Modality-Specific Refinement and Boundary Localization

D.5.1 Local Evidence Supplementation

To examine how local evidence should supplement dense reconstruction, we compare the designs in Table 13. All variants retain assignment-guided dense reconstruction. *w/o Local Residual* removes the residual branch; *Raw Residual Bypass* directly adds back original modality features without residual experts or allocation-conditioned gating; and *Shared Residual Expert* shares one

Table 12: Standard deviations are computed over the per-seed paired clean-versus-shifted drops.Average-mAP drops under controlled temporal misalignment (percentage points; lower is better). Negative values indicate slightly higher scores after shifting.

Model	± 0.5 s	± 1 s	± 2 s	± 4 s
LLAU	-0.10 ± 0.07	0.24 ± 0.14	0.51 ± 0.25	0.92 ± 0.28
w/o Offset	0.12 ± 0.10	0.56 ± 0.22	1.34 ± 0.37	2.11 ± 0.45
GLFK	-0.20 ± 0.23	0.43 ± 0.20	1.02 ± 0.33	1.55 ± 0.40

residual expert across modalities. *Residual w/o Assignment* retains modality-specific residual experts but removes final-assignment conditioning from residual gating and expert-output weighting, while preserving assignment information in the reconstruction branch. Full LLAU uses the final soft assignments $\mathbf{A}^{(R)}$ to connect dense reconstruction with modality-specific refinement through assignment-dependent residual weights.

Table 13: Comparison of local refinement designs.

Variant	Average-mAP (%)	Average-wMAP (%)
w/o Local Residual	15.90 ± 0.50	24.86 ± 0.38
Raw Residual Bypass	15.95 ± 0.89	25.16 ± 1.44
Shared Residual Expert	16.58 ± 0.74	25.39 ± 0.96
Residual w/o Assignment	16.75 ± 0.56	25.67 ± 0.63
Full LLAU	$\mathbf{17.47} \pm 0.53$	$\mathbf{26.41} \pm 0.36$

Full LLAU improves Average-mAP by 1.57 percentage points over the variant without local residual refinement. Directly adding back original features yields a mean Average-mAP gain of only 0.05 percentage points and remains 1.52 points below the full model. The shared-expert and assignment-unconditioned variants trail full LLAU by 0.89 and 0.72 points, respectively, with the same ordering in Average-wMAP. These results support modality-specific transformations and assignment conditioning for local evidence supplementation, rather than simply adding an original-feature bypass.

D.5.2 Boundary Error and Ground-Truth Coverage

All models use class-wise hard NMS with a tIoU threshold of 0.3, a minimum score of 0.001, and at most 2,000 segments per video. Boundary analysis further retains predictions with confidence scores of at least 0.05. Within each video and class, the Hungarian algorithm maximizes the total tIoU between predicted and ground-truth segments; assigned pairs with tIoU below 0.3 are then discarded.

Ground-truth (GT) coverage is the fraction of all GT segments successfully matched to predictions. Normalized boundary error is computed only on GT segments successfully matched by every compared model. For each such segment, we divide the mean absolute start- and end-boundary error by its GT duration, then average over this common set. Boundary accuracy is thus compared on the same GT segments, while coverage measures the extent of successful detection for each model.

Table 14: Effects of local refinement on normalized boundary error and GT coverage.

Model	Normalized boundary error $\downarrow$	GT coverage (%) $\uparrow$
Full LLAU	$\mathbf{0.238} \pm 0.009$	$\mathbf{85.1} \pm 0.8$
w/o Local Residual	0.276 ± 0.013	82.4 ± 1.3
Residual w/o Assignment	0.253 ± 0.010	84.0 ± 1.0
EMOE	0.260 ± 0.012	82.9 ± 1.2

Removing local residual refinement increases normalized boundary error by 0.038 and reduces GT coverage by 2.7 percentage points. Retaining residual experts without assignment conditioning partially closes this gap, while full LLAU further reduces boundary error by 0.015 and increases coverage by 1.1 percentage points. Compared with the external reference EMOE, full LLAU also achieves lower boundary error and higher coverage. These comparisons support assignment-conditioned residual refinement for boundary localization.

D.6 Hyperparameter Sensitivity

We examine the number of latent affective units per group, K, and the maximum modality offset, $\delta_{\max}$, in Table 15. Here, K specifies the number of units in the shared group and in each modality-specific group. The parameter $\delta_{\max}$ corresponds to `max_lag_steps` in the implementation and bounds the continuously predicted offsets as $\delta_{m,k} \in [-\delta_{\max}, \delta_{\max}]$; it does not denote a total window width. With a feature step of 0.5 seconds, bounds of 2, 4, and 8 steps allow maximum offsets of ± 1, ± 2, and ± 4 seconds, respectively. When varying this bound, the cross-modal support search radius remains fixed at $\tau = 4$, and the unit-width configuration is unchanged.

Table 15: Hyperparameter sensitivity. Default settings are shown in bold.

Hyperparameter	Setting	Average-mAP (%)	Average-wMAP (%)
Units per group K	8	16.67 ± 0.73	25.35 ± 0.58
	16	$\mathbf{17.47} \pm 0.53$	$\mathbf{26.41} \pm 0.36$
	32	16.04 ± 1.47	25.10 ± 1.44
Offset bound $\delta_{\max}$ (steps)	2	16.97 ± 0.95	26.15 ± 0.62
	4	$\mathbf{17.47} \pm 0.53$	$\mathbf{26.41} \pm 0.36$
	8	16.27 ± 1.25	25.54 ± 1.06

The default configuration, $K = 16$ and $\delta_{\max} = 4$, achieves the highest mean performance among the evaluated settings. Increasing either the unit count or the offset bound provides no additional gains. These results indicate that more units or a larger offset range do not necessarily improve localization performance, supporting the current compact configuration.

Weight of unit auxiliary supervision. We further examine the coefficient λ_{unit} of the unit auxiliary loss. As shown in Fig. 7, the preset default $\lambda_{\text{unit}} = 0.1$ achieves the highest mean Average-mAP and Average-wMAP among the evaluated settings. Setting the weight to zero reduces Average-mAP from 17.47 to 16.19, while larger weights provide no additional benefit.

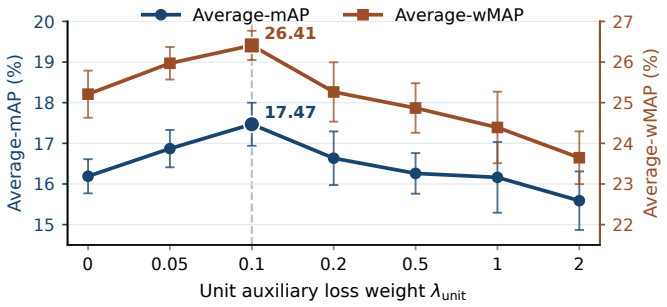

Figure 7: Sensitivity to the unit auxiliary loss weight. Mean Average-mAP (left axis) and Average-wMAP (right axis) are shown together; error bars show one standard deviation across seeds, and both axes have the same percentage-point scale. The dashed line marks the preset default $\lambda_{\text{unit}} = 0.1$; evaluated weights are equally spaced along the horizontal axis for readability.

AI USE STATEMENT

We used large language models solely for language polishing, including improving grammar, clarity, and readability. All AI-assisted edits were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. The authors take full responsibility for the final content of this paper.