PSYCHE: MODELING CONVERSATIONAL EMOTION AS A CONTINUOUS TRAJECTORY VIA CAUSAL SOURCE DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Emotion Recognition in Conversation (MERC) faces a fundamental challenge: textual, acoustic, and visual expressions are sparse, regulation-filtered observations of internal emotional states that evolve continuously. Genuine calm and suppressed disappointment can exhibit nearly identical surface signals yet arise from opposite trajectories. Existing methods harness rich conversational context and multimodal cues, yet model emotion as a discrete per-utterance label rather than an evolving state. Shifts are consequently detected by auxiliary binary heads, while causes are attributed post hoc by separate modules, so that *how* emotion evolves and *why* it changes are never decoded by a single process. We propose PSYCHE, which models each speaker's emotion as a latent trajectory governed by a Neural Jump ODE: Causal Source Decomposition (CSD) structures the driving force for continuous inter-utterance evolution, while speech acts trigger discrete updates. CSD decomposes this driving force into a self-cause driver with state-dependent decay and an other-cause driver attending to the interlocutor's triggering utterances, fused by a per-dimension gate. The self-/other-cause balance and trigger identification are computed by the forward pass itself, requiring no post-hoc explainer or inference-time language model. On IEMOCAP and MELD, PSYCHE achieves the best weighted F1 among the baselines we re-evaluate under identical multimodal features, with offline commonsense cues as an added input. Dynamics analyses show that the model lowers the self-cause ratio when emotion turns and reduces confusion precisely where surface signals mask internal states, indicating that *how* and *why* are not two answers but one mechanism.

1 INTRODUCTION

A person excitedly shares, "I got the job!" A friend replies indifferently, "Mm, that's nice." After a brief silence, the first person says quietly, "Forget it, it's nothing." A classification model labels this last utterance neutral, and an attribution model identifies which earlier utterance it relies on. Yet someone who has followed the whole conversation reads something different. The indifferent reply shifts the speaker from excitement toward disappointment, and the silence that follows lets the shift deepen before it surfaces as a quiet "forget it." A classifier sees a neutral label, yet a human observer traces a continuous arc from excitement to disappointment. The difference lies in what happens between words: the emotional trajectory that surface signals can only partially reveal.

Gross's emotion-regulation theory (Gross, 1998) motivates this distinction: internal affect evolves with inertia, perturbation, and regulation, whereas words, prosody, and expressions provide only sparse, filtered observations. Figure 1 visualizes this for our opening story: the trajectory evolves between turns under self-inertia drift and other-cause perturbation, yet surfaces as a flat "neutral" that conceals the accumulating emotional shift. Predicting emotion from surface features alone cannot bridge this gap.

Existing methods follow two paths. One focuses on classification, progressing from recurrent networks (Hazarika et al., 2018; Majumder et al., 2019) to graph neural networks (Ghosal et al., 2019; Shen et al., 2021) and cross-modal fusion (He et al., 2026), with recent work adding aux-

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.

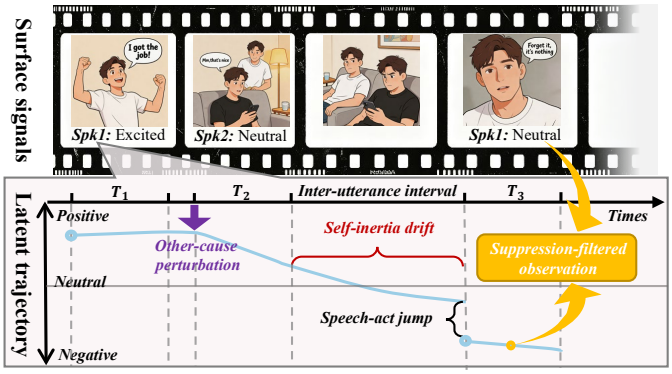


Figure 1: The "good news met with indifference" case, illustrated as a comic strip. The latent trajectory (blue) evolves continuously between utterances, driven by self-inertia drift and other-cause perturbation, with discrete speech-act jumps.

iliary heads to flag whether emotion *shifts* between adjacent utterances (Li et al., 2024; 2025). The other path seeks to explain emotion, typically by prompting large language models to reason about inter-utterance causes (Jing et al., 2026), retrieve emotion cues (Li et al., 2026), or decompose cognition through multi-agent LLM systems (Nguyen et al., 2026). Both advance the task yet share the premise that emotion remains a discrete per-utterance label. The richer the context a classifier consumes, the better it predicts that label, but a label is a snapshot carrying no dynamics from which a shift can be derived. Shifts are consequently detected as binary events rather than outcomes of an evolving state, while explanations are produced post hoc by an external language model rather than by the mechanism that predicts the emotion. What is still missing is a single process in which *how* an emotion evolves and *why* it changes arise together, readable without a separately deployed shift detector or an online LLM at inference.

This motivates our central question: can a latent state trajectory evolving continuously with the conversation let *how* emotion changes and *why* it changes be derived from one shared process, rather than two unrelated modules? We approach this question from three angles. (i) **Continuity.** Modeling emotion as a trajectory rather than a sequence of discrete labels lets the evolving latent state summarize conversational history, so that genuine calm and suppressed disappointment become distinguishable by the path that led to them, even when their surface signals are nearly identical. (ii) **Single-pass attribution.** When the same mechanism that drives state evolution also governs what the decoder reads, the signal of *why* a transition occurred is available from the forward pass itself, without a post-hoc explainer or an inference-time language model. (iii) **Independent verifiability.** Because prediction, transition direction, and change attribution share one source, they corroborate one another and can each be examined against separate external annotations. We target the observable: transitions derived from dynamics, and attribution decoded from the forward pass.

We realize these three design angles in **PSYCHE** (from the Greek $\psi\upsilon\chi\acute{\eta}$, the inner life beneath outward expression). For continuity, a causal graph encoder fuses multimodal signals within each turn and propagates context across turns under a strict temporal order, routing information into separate self- and other-speaker streams. A Neural Jump ODE then integrates the causal control as the sole driver of continuous state evolution between utterances, with discrete speech-act jumps. For single-pass attribution, Causal Source Decomposition (CSD), the technical core, structures the ODE's driving force into a self-cause driver and an other-cause driver, fused by a per-dimension gate. The self-cause driver captures emotional inertia with state-dependent decay over longer gaps, while the other-cause driver attends to the interlocutor's intervening utterances via learned attention. Because the ODE alone governs evolution while CSD only defines its input, the two drivers remain disentangled, and the self-cause ratio and trigger attention are read off as forward-pass by-products. A decoupled decoder reads the label from the pre-speech state and the two drivers separately, so that the causal source of a shift, not merely its magnitude, determines the prediction. For independent verifiability, the self-cause ratio and trigger attention are validated against human-annotated references without any component trained for attribution.

In a nutshell, our contributions are as follows:

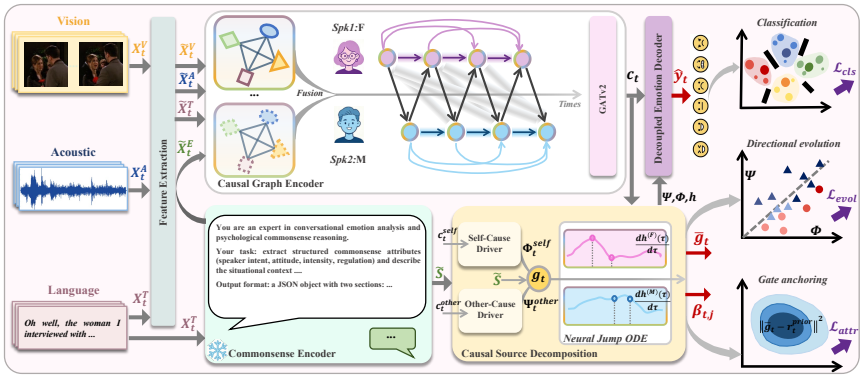


Figure 2: Architecture of **PSYCHE**. *Red arrows*: the three outputs (prediction $\hat{\mathbf{y}}_t$, self-cause ratio $\bar{g}_t$, trigger attention $\{\beta_{t,j}\}$). *Purple arrows*: the three training objectives ($\mathcal{L}_{\text{cls}}$, $\mathcal{L}_{\text{evol}}$, $\mathcal{L}_{\text{attr}}$).

- We recast conversational emotion recognition from classifying surface signals to modeling a continuously evolving latent trajectory, so that transitions are derived from the trajectory itself rather than detected by a separate shift detector at inference.
- We propose CSD, which, to our knowledge, is the first to decompose a Neural Jump ODE's sole causal control into explicit self- and other-cause drivers fused by a per-dimension gate. Prior ODE-based emotion models drive the vector field with raw or graph-diffused observations, offering no built-in attribution.
- CSD yields intrinsic interpretability: the other-cause attention localizes the trigger without attribution supervision, and the per-dimension gate provides a self-cause ratio anchored by a weak offline prior. Both are available at inference, requiring no post-hoc explainer or inference-time language model.

2 Related Work

MERC. Conversational emotion recognition has progressed from recurrent context modeling (Hazarika et al., 2018; Majumder et al., 2019) to speaker- and temporal-relation graphs (Ghosal et al., 2019; Shen et al., 2021) and cross-modal fusion (He et al., 2026). Recent work addresses emotion shifts: CFN-ESA (Li et al., 2024) and GraphSmile (Li et al., 2025) add auxiliary shift-detection heads, while EMART (Seong et al., 2026) shapes the label space via an emotion-wheel contrastive loss. All three still model emotion as a per-utterance classification target, so that shifts remain binary predictions rather than outcomes of an evolving state. PSYCHE instead derives transitions from continuous latent dynamics.

Interpretable and Causal ERC. A second line seeks to explain emotion. ECERC (Zhang & Tan, 2025) separates emotional evidence from contextual causes, yielding cause-specific attention weights, and UniMEEC (Hu et al., 2024) unifies multimodal emotion recognition with emotion-cause analysis. Several methods instead delegate reasoning to LLMs (Jing et al., 2026; Li et al., 2026; Nguyen et al., 2026). These treat attribution as a primary objective but produce it post-hoc, and for LLM-based approaches this requires an inference-time language model. PSYCHE differs in two respects. First, its attribution is a by-product of the same forward pass that predicts the label, computed by learned parameters without any inference-time language model, and its trigger attention is unsupervised rather than separately trained. Second, unlike ECERC's static per-utterance attention, PSYCHE's self/other separation drives a continuous-time trajectory whose self-cause contribution decays with elapsed time.

Neural ODEs for Emotion Dynamics. Continuous-time affect modeling via neural ODEs is a recent direction: DCDNODE (Wu et al., 2024) predicts continuous valence/arousal for a single speaker, and DGODE (Shou et al., 2025) evolves multi-party node representations through a graph ODE. In both, the vector field is driven by raw or graph-aggregated observations without decomposing the speaker's inertia from interlocutor influence, and the dynamics carry no built-in attribution. CSD addresses this gap by decomposing the sole causal control of a Neural Jump ODE into ex-

plicit self- and other-cause components fused by a per-dimension gate, so that the self-cause ratio and trigger attention are read directly off the decomposition rather than produced by a post-hoc explainer.

3 METHOD

3.1 OVERVIEW

Given a T-utterance conversation with speaker set $\mathcal{P}$ and raw multimodal inputs $\mathbf{X}_t^m$ ($m \in \{\mathrm{T}, \mathrm{A}, \mathrm{V}\}$), PSYCHE jointly predicts per-utterance emotion labels, a self-cause ratio $\bar{g}_t$, and trigger attention $\{\beta_{t,j}\}$ for each turn. Five modules constitute the pipeline (Fig. 2). First, a **Feature Extraction Module** encodes $\mathbf{X}_t^m$ into fixed-length representations. Second, a **Commonsense Encoder** supplies the situational grounding that perceptual signals alone lack. It extracts an entity-level cue $\tilde{\mathbf{x}}_t^{\mathrm{E}}$ per utterance and a scene-level cue $\tilde{\mathbf{s}}$ per dialogue along psychologically salient dimensions that reflect the inner-state and regulation factors of Gross's account. These cues give the trajectory a stable reference of *what the conversation is about*, helping it distinguish, for instance, composure at a funeral from composure that masks disappointment. The scene-level cue $\tilde{\mathbf{s}}$ encodes only the static situational background available at the onset of the dialogue (speaker identities, physical setting, power dynamics); it does not encode any later dialogue content, so it introduces no future information into the per-turn state. The entity-level cue $\tilde{\mathbf{x}}_t^{\mathrm{E}}$ likewise captures the speaker's stable persona and attitude rather than the outcome of future turns. Both cues are extracted once by an offline LLM as static, frozen features and encoded by a frozen RoBERTa-large, so the temporal causality of the graph propagation and ODE evolution—which use only these fixed inputs—is not affected. Consistent with the knowledge-aware MERC line (Ghosal et al., 2020; Ren et al., 2025), they operate as an internal module, not an external service, so no LLM call occurs during training or inference. Third, a **Causal Graph Encoder** fuses the four modality cues within each turn and propagates context across turns under strict temporal order, routing the result into a self-stream $\mathbf{c}_t^{\mathrm{self}}$ and other-stream $\mathbf{c}_t^{\mathrm{other}}$. Fourth, **Causal Source Decomposition** turns the two streams into a self-cause driver $\mathbf{\Phi}_t^{\mathrm{self}}$ and an other-cause driver $\mathbf{\Psi}_t^{\mathrm{other}}$, fuses them via a per-dimension gate into a single causal control $\mathbf{m}_t$, and lets a Neural Jump ODE integrate $\mathbf{m}_t$ as the sole driver of continuous evolution between turns, with discrete speech-act jumps at each turn. Fifth, a **Decoupled Emotion Decoder** reads the label from the pre-speech state and the two drivers fed separately. Because CSD constructs the two drivers separately before fusion, $\mathbf{\Phi}_t^{\mathrm{self}}$ and $\mathbf{\Psi}_t^{\mathrm{other}}$ remain explicitly accessible in the forward pass.

3.2 CAUSAL GRAPH ENCODER

The encoder resolves cross-modal conflicts within each utterance and propagates context across turns under a temporal causality constraint, routing signals into separate self- and other-pathways so that CSD can drive the trajectory from distinct sources.

3.2.1 INTRA-TURN MULTIMODAL FUSION.

Each utterance t carries text $\tilde{\mathbf{x}}_t^{\mathrm{T}}$, audio $\tilde{\mathbf{x}}_t^{\mathrm{A}}$, visual $\tilde{\mathbf{x}}_t^{\mathrm{V}}$, and the entity cue $\tilde{\mathbf{x}}_t^{\mathrm{E}}$, each projected to $\mathbb{R}^d$ by a linear layer with ELU and dropout. A modality reliability gate $\boldsymbol{\rho}_t = \sigma(\mathbf{W}_\rho[\tilde{\mathbf{x}}_t^{\mathrm{T}} \| \tilde{\mathbf{x}}_t^{\mathrm{A}} \| \tilde{\mathbf{x}}_t^{\mathrm{V}}] + \mathbf{b}_\rho) \in (0,1)^3$ scales each perceptual modality element-wise, e.g. suppressing an unreliable text channel when the words claim composure but the voice trembles. The entity cue is not gated, as it is not a sensor modality. The four nodes form a fully connected graph with a shared learnable edge embedding. Graph attention with residual connections fuses the four nodes, and mean pooling yields $\mathbf{u}_t^{(0)}$.

3.2.2 INTER-TURN TEMPORAL PROPAGATION.

Propagation runs on a strictly temporally admissible directed graph: an edge $k \to t$ exists only when $k < t$, with a causal mask enforcing $\alpha_{t \leftarrow k} = 0$ for $k \geq t$. Edges carry type embeddings, self if $\mathrm{spk}(k) = \mathrm{spk}(t)$ and other otherwise. This temporal constraint reflects time-admissibility of causal effect rather than causal discovery. We stack L=2 layers of GATv2 (Brody et al., 2022), whose dynamic attention conditions on both source and target, supporting the self-vs-other distinction. For

target t,

$$\mathbf{u}_t^{(\ell)} = \mathrm{GELU}\left(\sum_{k<t} \alpha_{t\leftarrow k}^{(\ell)} \cdot \mathbf{W}_v^{(\ell)} \mathbf{u}_k^{(\ell-1)}\right) + \mathbf{u}_t^{(\ell-1)}, \tag{1}$$

with $\mathbf{c}_t = \mathbf{u}_t^{(L)}$. At the final layer, attention is renormalized *within* each speaker subset to yield the self-stream $\mathbf{c}_t^{\mathrm{self}}$ (same-speaker) and other-stream $\mathbf{c}_t^{\mathrm{other}}$ (cross-speaker). Without this separation, CSD would receive one blended signal and could not drive the trajectory independently from the two sources.

3.3 Causal Source Decomposition (CSD)

CSD takes the two streams $\mathbf{c}_t^{\mathrm{self}}$ and $\mathbf{c}_t^{\mathrm{other}}$, computes a self-cause and an other-cause driver, fuses them via a per-dimension gate, and feeds the result as the sole input to a Neural Jump ODE. Because the ODE is the single channel of state evolution and CSD merely defines its driving force, the two causes remain disentangled rather than competing over the state update.

3.3.1 Self-Cause Driver: State-Dependent Inertia.

The self-cause driver models the speaker's *emotional inertia* (Kuppens et al., 2010), the tendency of an affective state to persist and decay. Let $\mathbf{h}_{\mathrm{prev}}^{+} = \mathbf{h}^{(p)}(t_{\mathrm{prev}}^{+})$ be the state after speaker p's most recent utterance at $t_{\mathrm{prev}} = \mathrm{prev}(p,t)$ (for a first utterance, $\mathbf{h}^{(p)}(t_0) = \mathbf{W}_{\mathrm{init}}\mathbf{c}_{\mathrm{first}(p)}$), and $\delta = (t - t_{\mathrm{prev}})/T \in (0,1]$ the normalized interval. The decay rate is state-dependent: $\gamma_t = \mathrm{softplus}(\mathbf{w}_\gamma^\top \mathbf{h}_{\mathrm{prev}}^{+} + b_\gamma)$, so that high-arousal states resist decay while low-arousal ones fade faster. Then

$$\mathbf{\Phi}_t^{\mathrm{self}} = \mathbf{W}_{\mathrm{self}} \cdot \tanh(\mathbf{h}_{\mathrm{prev}}^{+}) \cdot \exp(-\gamma_t \cdot \delta), \tag{2}$$

where $\exp(-\gamma_t\delta)$ attenuates influence over longer gaps and $\tanh$ bounds the contribution.

3.3.2 Other-Cause Driver: Interlocutor Perturbation.

The other-cause driver models how the interlocutor's words perturb the trajectory. Over the intervening partner utterances $\mathcal{O}(p,t) = \{j : t_{\mathrm{prev}} < j < t \wedge \mathrm{spk}(j) \neq p\}$, a bilinear attention with learnable temperature τ scores each candidate given the prior state:

$$\beta_{t,j} = \frac{\exp(s_{t,j}/\tau)}{\sum_{j' \in \mathcal{O}} \exp(s_{t,j'}/\tau)}, \quad s_{t,j} = {\mathbf{h}_{\mathrm{prev}}^{+}}^{\top} \mathbf{W}_\beta \, \mathbf{c}_j^{\mathrm{other}}. \tag{3}$$

The driver aggregates the attended candidates:

$$\mathbf{\Psi}_t^{\mathrm{other}} = \sum\nolimits_{j \in \mathcal{O}(p,t)} \beta_{t,j} \, \mathbf{W}_{\mathrm{other}} \mathbf{c}_j^{\mathrm{other}}. \tag{4}$$

The utterance with the highest attention is the *trigger*. Because $\{\beta_{t,j}\}$ receives *no* attribution supervision, it is determined purely by fitting the trajectory. It is evaluated against offline reference annotations on a human-verified subset.

3.3.3 Per-Dimension Gating and Fused Causal Control.

The two drivers fuse through a per-dimension gate that also receives a learnable Fourier time encoding $\mathbf{r}(\delta)$ and the scene embedding $\tilde{\mathbf{s}}$:

$$\mathbf{g}_t = \sigma(\mathbf{W}_g[\mathbf{\Phi}_t^{\mathrm{self}} \| \mathbf{\Psi}_t^{\mathrm{other}} \| \mathbf{r}(\delta) \| \tilde{\mathbf{s}}] + \mathbf{b}_g) \in (0,1)^d, \tag{5}$$

$$\mathbf{m}_t = \mathbf{g}_t \odot \mathbf{\Phi}_t^{\mathrm{self}} + (1 - \mathbf{g}_t) \odot \mathbf{\Psi}_t^{\mathrm{other}}. \tag{6}$$

A per-dimension gate preserves which dimensions are self-driven and which are other-driven, while a scalar gate would average this structure away. The scalar mean $\bar{g}_t = \frac{1}{d}\sum_i g_{t,i}$ is the self-cause ratio: near 1 the control is self-dominated, near 0 other-dominated, a by-product analyzed in the Interpretability and Dynamics Analysis section.

Datasets	LR	BS	EP	DR	λ_{attr}	λ_{evol}	λ_{dir}	λ_{gate}	d
IEMOCAP	1e-4	4	40	0.5	0.05	0.15	0.05	0.3	512
MELD	5e-5	8	25	0.3	0.03	0.10	0.05	0.3	256

Table 1: Hyperparameters: learning rate (LR), batch size (BS), epochs (EP), dropout (DR), loss weights, and hidden-dimension (d).

Methods	**IEMOCAP**								**MELD**	
	Hap.	*Sad*	*Neu.*	*Ang.*	*Exc.*	*Fru.*	***WA***(↑)	***F1***(↑)	***WA***(↑)	***F1***(↑)
Bc-LSTM	33.50	73.02	54.69	62.39	61.62	59.74	59.09	58.84	59.54	57.22
DialogueGCN	31.76	80.35	54.12	61.54	72.43	59.15	62.17	61.43	59.16	56.78
DialogueRNN	33.80	83.47	55.26	62.35	73.91	60.00	64.02	62.91	56.13	57.39
MMGCN	42.11	77.97	64.50	63.01	70.15	63.66	65.25	65.23	65.98	64.78
MM-DFN	41.86	80.51	70.15	64.33	72.10	67.25	68.45	68.28	66.40	65.21
GCCL	50.37	78.01	70.47	68.53	74.63	68.07	70.14	69.83	65.89	65.29
CFN-ESA	53.29	80.72	69.69	**68.73**	74.60	67.29	70.06	70.14	67.28	66.04
ECERC	54.85	79.19	71.04	66.48	74.86	67.55	69.81	70.24	67.51	66.39
GraphSmile	56.59	**81.91**	68.98	67.57	76.49	70.06	71.16	71.32	67.16	65.56
EMART	53.55	80.45	66.49	61.94	72.95	65.92	68.06	68.02	66.63	65.30
PSYCHE	**60.58**	81.15	**72.00**	67.96	**77.41**	**70.16**	**72.58**	**72.51**	**68.62**	**67.55**

Table 2: Comparison with state-of-the-art on IEMOCAP and MELD. Baselines are retrained on our splits from released code. We report per-class F1, Weighted Accuracy (***WA***), and Weighted F1 (***F1***) in %; **best** and second-best per column highlighted.

3.3.4 Neural Jump ODE: Continuous Evolution and Speech-Act Jump.

The causal control $\mathbf{m}_t$ conditions the Neural Jump ODE (Jia & Benson, 2019) vector field rather than additively updating the state. Collecting the conditioning terms as $\mathbf{z}_t = [\mathbf{m}_t \| \mathbf{c}^{\text{self}}_{t_{\text{prev}}} \| \mathbf{c}^{\text{other}}_{t_{\text{prev}}} \| \tilde{\mathbf{s}} \| \mathbf{r}(\delta)]$, the state evolves between utterances as

$$\frac{d\mathbf{h}^{(p)}(\tau)}{d\tau} = f_\theta\big(\mathbf{h}^{(p)}(\tau),\, \mathbf{z}_t\big), \quad \tau \in [0, \delta], \tag{7}$$

with f_θ a two-layer GELU MLP. The two context streams in $\mathbf{z}_t$ are taken from the previous utterance rather than the current turn, preventing the label-bearing turn t from leaking into the pre-speech state. Since $\mathbf{\Psi}^{\text{other}}_t$ aggregates all partner utterances in the interval into a single control, evolution is *interval-conditioned* rather than event-wise causal, yet the endpoint $\mathbf{h}(t^-)$ depends only on information before turn t. We use a fixed 4-step RK4 solver, chosen for its static computation graph, reproducibility, and direct backpropagation, to obtain $\mathbf{h}(t^-) = \text{RK4}(f_\theta, \mathbf{h}^+_{\text{prev}}, \mathbf{z}_t, \delta)$, the state just before speaking. At each speech act, a *speech-act jump* (inspired by self-perception theory (Bem, 1972), where voicing an emotion reshapes it) produces the post-speech state

$$\mathbf{h}^{(p)}(t^+) = \mathbf{h}(t^-) + \mathbf{j}_\phi\big(\mathbf{h}(t^-), \mathbf{c}^{\text{self}}_t\big) \cdot \mathbb{I}[\text{spk}(t) = p], \tag{8}$$

only the current speaker jumps while others evolve continuously. The jump is conditioned solely on the pre-speech state and self-stream, preventing train–inference mismatch.

3.4 Decoupled Emotion Decoder

The decoder predicts the emotion label from the pre-speech state and the two drivers fed separately:

$$\hat{\mathbf{y}}_t = g_\psi\big(\mathbf{c}_t,\, \mathbf{h}^{(p)}(t^-),\, \bar{\mathbf{h}}^{(p)},\, \mathbf{\Phi}^{\text{self}}_t,\, \mathbf{\Psi}^{\text{other}}_t\big), \tag{9}$$

where $\bar{\mathbf{h}}^{(p)}$ is the running mean of speaker p's historical post-speech states (a strictly historical baseline) and g_ψ is a two-layer GELU MLP. Feeding $\mathbf{\Phi}^{\text{self}}_t$ and $\mathbf{\Psi}^{\text{other}}_t$ separately exposes source-specific structure and discourages reliance on the magnitude of the fused control $\mathbf{m}_t$ alone.

3.5 TRAINING OBJECTIVE

The total objective combines three terms:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{evol}}\mathcal{L}_{\text{evol}} + \lambda_{\text{attr}}\mathcal{L}_{\text{attr}}. \tag{10}$$

Prediction. $\mathcal{L}_{\text{cls}}$ is a class-weighted cross-entropy with label smoothing, the only term that reads the emotion label directly.

Evolution. This objective shapes *how* the drivers move without accessing the state, and consists of three sub-terms. (i) A training-only binary head $\hat{\tau}_t = \sigma(\mathbf{w}_\tau^\top[\mathbf{\Phi}_t^{\text{self}} \| \mathbf{\Psi}_t^{\text{other}}] + b_\tau)$ discarded at inference, regularizes whether the emotion changes ($\tau_t = \mathbb{I}[y_t \neq y_{\text{prev}}]$), yielding $\mathcal{L}_{\text{evol}}^{\text{bin}}$. Using the concatenated pair rather than a scalar, this head probes whether $\mathbf{\Phi}$ and $\mathbf{\Psi}$ are structurally separable. (ii) A directional term

$$\mathcal{L}_{\text{evol}}^{\text{dir}} = 1 - \cos\big(\mathbf{\Psi}_t^{\text{other}},\ \mathbf{e}_{y_t} - \mathbf{e}_{y_{\text{prev}}}\big), \tag{11}$$

aligns *only* the other-cause driver with the semantic direction of change, breaking its symmetry with $\mathbf{\Phi}$. (iii) A gate-margin term penalizes shifts whose self-cause ratio is not sufficiently below that of stays:

$$\mathcal{L}_{\text{gate}} = \max\big(0,\ m - (\bar{g}^{\text{stay}} - \bar{g}^{\text{shift}})\big), m = 0.15. \tag{12}$$

Together $\mathcal{L}_{\text{evol}} = \mathcal{L}_{\text{evol}}^{\text{bin}} + \lambda_{\text{dir}}\mathcal{L}_{\text{evol}}^{\text{dir}} + \lambda_{\text{gate}}\mathcal{L}_{\text{gate}}$.

Attribution. Without supervision the gate may permute its self/other roles, so to fix their *semantics* we regress the gate mean to an offline prior via $\mathcal{L}_{\text{attr}} = \|\bar{g}_t - r_t^{\text{prior}}\|^2$, where $r_t^{\text{prior}} \in [0, 1]$ is a per-utterance self-cause ratio distilled once offline by an LLM annotator, distinct from the commonsense encoder, and used only as a light mean-anchoring target. This prior anchors which side of the decomposition means self and which means other without dictating the gate's per-utterance values, leaving the gate driven by $\mathbf{\Phi}$ and $\mathbf{\Psi}$ themselves. Likewise, $\mathcal{L}_{\text{gate}}$ only enforces the *relative* order of $\bar{g}^{\text{stay}}$ over $\bar{g}^{\text{shift}}$, not their per-utterance magnitudes. The trigger attention $\{\beta_{t,j}\}$ is left fully unsupervised, so trigger localization is not directly supervised.

More method details are provided in the Appendix.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We evaluate on **IEMOCAP** (Busso et al., 2008) and **MELD** (Poria et al., 2018), the standard benchmarks for MERC. Following prior work (Joshi et al., 2022; Li et al., 2025), we adopt the same dataset splits and feature preprocessing. Text features are obtained from RoBERTa-large (Liu et al., 2019), visual features from DenseNet (Huang et al., 2017), and acoustic features from OpenSMILE (Eyben et al., 2010).

Implementation. All experiments are run on an NVIDIA RTX 3090 GPU. The model is optimized with AdamW under a class-balanced cross-entropy loss with label smoothing (0.1). Dataset-specific hyperparameters are in Table 1. All results are from the single best validation checkpoint, and training stability is verified under three random seeds. More implementation details are provided in the Appendix.

Baselines. We compare against methods spanning four paradigms: recurrent (Bc-LSTM (Poria et al., 2017), DialogueRNN (Majumder et al., 2019)), graph-based (DialogueGCN (Ghosal et al., 2019), MMGCN (Hu et al., 2021), MM-DFN (Hu et al., 2022), GCCL (Dai et al., 2024), CFN-ESA (Li et al., 2024)), emotion-shift-aware (GraphSmile (Li et al., 2025), ECERC (Zhang & Tan, 2025), EMART (Seong et al., 2026)), and LLM-based attribution (CMTD (Nguyen et al., 2026), PRC-Emo (Li et al., 2026), Causal-ERC (Jing et al., 2026)).

4.2 COMPARISON WITH STATE-OF-THE-ART

Table 2 reports the main results. For fairness, each available baseline is retrained from public code using the same splits, raw multimodal features, and evaluation protocol; PSYCHE additionally uses the offline commonsense cues of Table 4. PSYCHE achieves the best weighted F1 on both datasets,

Methods	*Inf. LLM*	*Intrinsic Attr.*	*WA*	*W-F1*	*ΔW-F1*
CMTD†	✓	×	64.23	64.17	▼8.34
PRC-Emo†	✓	×	71.03	71.95	▼0.56
Causal-ERC†	✓	×	71.66	71.89	▼0.62
PSYCHE	×	✓	**72.58**	**72.51**	–

Table 3: Comparison with LLM-based attribution methods on IEMOCAP. † indicates results reported in original papers. *Inf. LLM*: inference-time language model required. *Intrinsic Attr.*: attribution via forward pass.

Variants	*WA*	*W-F1*	*ΔW-F1*
Core Architecture			
-w/o CSD	67.47	67.54	▼4.97
-w/o self-/other split	68.78	68.60	▼3.91
-Scalar gate	71.78	71.56	▼0.95
-Fixed γ	72.01	71.72	▼0.79
ODE Architecture			
-w/o ODE	70.45	70.21	▼2.30
-w/o speech-act jump	71.32	71.09	▼1.42
-w/o Fourier time encoding	71.67	71.44	▼1.07
Loss Functions			
-w/o $\mathcal{L}_{\text{attr}}$	71.05	70.73	▼1.78
-w/o $\mathcal{L}_{\text{evol}}$	69.31	69.38	▼3.13
-w/o $\mathcal{L}_{\text{dir}}$	70.05	69.99	▼2.52
-w/o $\mathcal{L}_{\text{gate}}$	72.14	71.92	▼0.59
Commonsense Features			
-w/o Entity commonsense	71.56	71.30	▼1.21
-w/o Scene commonsense	72.05	71.82	▼0.69
-w/o Both commonsense	71.02	70.74	▼1.77
PSYCHE	**72.58**	**72.51**	–

Table 4: Ablation results on IEMOCAP.

surpassing GraphSmile and CFN-ESA by 1.19% and 2.37% on IEMOCAP, and by 1.16% over ECERC on MELD. On IEMOCAP it ranks among the top two in every emotion category (full per-class MELD results are reported in the Appendix). PSYCHE's result is stable across seeds $\{22, 1246, 3407\}$, with mean$\pm$std W-F1 72.06 ± 0.34 on IEMOCAP and 67.21 ± 0.43 on MELD (full table in the Appendix). These findings support the value of formulating emotion recognition as continuous trajectory modeling.

Table 3 compares PSYCHE with three LLM-based attribution methods on IEMOCAP (results taken from the original papers, as no public code is available). PSYCHE achieves the best weighted F1 among them, while producing attribution intrinsically from its forward pass without a separate inference-time language model.

4.3 Ablation Studies

We ablate along four axes. Table 4 reports the results. Every removal degrades performance. Removing CSD costs 4.97 points, the largest drop by a wide margin, confirming causal decomposition as the primary architectural contribution. The self-/other-stream separation (3.91) and the evolution loss $\mathcal{L}_{\text{evol}}$ (3.13) form the second tier, showing that the decomposition requires both architecturally separated speaker streams and joint supervision of transition dynamics. The gate-anchoring prior $\mathcal{L}_{\text{attr}}$ contributes a moderate 1.78, consistent with its role as an orientation fix rather than a value supervisor. The contrastive gate margin costs only 0.59, so the gate is driven by the drivers themselves. Removing both commonsense cues costs 1.77 points, indicating that the offline cues provide a non-trivial complementary signal on top of the trajectory and CSD. These findings support the design as a coherent whole.

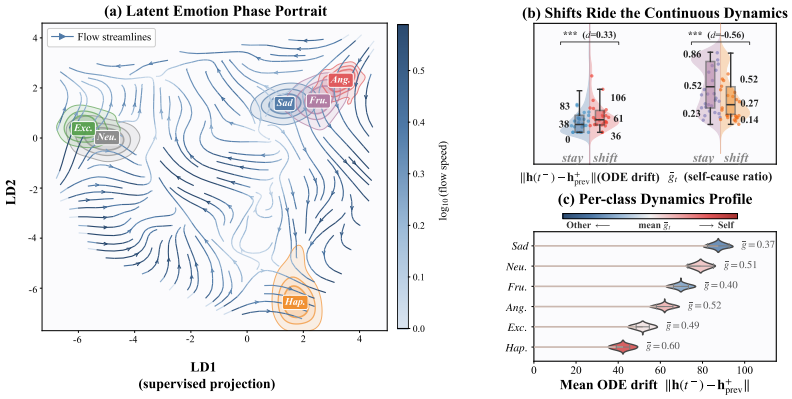


Figure 3: Latent trajectory dynamics on IEMOCAP. **(a)** Phase portrait (LDA projection): gradient-shaded emotion basins and flow streamlines confirm non-degenerate latent dynamics. **(b)** Split-bean plots of ODE drift and self-cause ratio $\bar{g}_t$ at stays (left) vs. shifts (right), with overlaid box statistics. **(c)** Per-class dynamics: lollipop stems show mean drift with SEM bars, head colors indicate $\bar{g}_t$ (blue other-driven, red self-driven), and mini-violins show per-dimension gate heterogeneity.

4.4 Interpretability and Dynamics Analysis

4.4.1 Latent Trajectory Dynamics.

Figure 3 validates that the Neural Jump ODE yields non-degenerate, structured dynamics. Panel (a) shows organized emotion basins and coherent flow streamlines in a supervised 2-D projection. Panel (b) provides quantitative evidence: ODE drift is significantly larger at emotion shifts than at stays (80.3 vs. 56.1, $p{=}7{\times}10^{-11}$, Cohen's $d{=}0.33$), and the self-cause ratio $\bar{g}_t$ drops from 0.53 to 0.36 at shifts ($p{=}7{\times}10^{-28}$, $d{=}0.56$); since drift magnitude receives no direct supervision, its stronger presence at transitions indicates the ODE drives state evolution rather than merely encoding the speech-act jump. Panel (c) shows happiness is self-driven ($\bar{g}{=}0.60$) while sadness and frustration are other-driven ($\bar{g}{\approx}0.38$), consistent with affective inertia theory (Kuppens et al., 2010; Butler, 2011). This trajectory also reduces emotion-to-neutral leakage from 127 to 80 on IEMOCAP (-37.0%), separating suppressed negative states from genuine calm whose surface signals are nearly identical.

Trigger localization. We evaluate the other-cause attention $\{\beta_{t,j}\}$ against non-learning baselines on RECCON's (Poria et al., 2021) multi-candidate transitions (visualizations in the Appendix). The unsupervised β achieves an NDCG of 0.83 and MAP of 0.77, outperforming the Uniform baseline (NDCG 0.76, MAP 0.67); where the trigger is not the most recent partner utterance it recovers the true trigger far above chance (85.7% vs. 50.0%). A large-scale offline-annotated proxy evaluation is also reported in the Appendix.

Self-cause ratio semantics. Grouping $\bar{g}_t$ by RECCON causal type (see the Appendix), the self-cause ratio is $0.64{\pm}0.29$ for self-contagion versus $0.43{\pm}0.28$ for inter-personal ($d{=}0.75$, AUROC$=0.71$), while hybrid and no-context lie in between. The monotonic gradient indicates the gate reflects the annotated causal structure; this is not circular since the prior is a continuous LLM ratio whereas RECCON types are categorical labels from an independent scheme.

5 Conclusion

PSYCHE models emotion as a continuous-time latent trajectory driven by Causal Source Decomposition, decomposing a Neural Jump ODE's control into self-cause and other-cause drivers fused per dimension. PSYCHE achieves 72.51%/67.55% W-F1 on IEMOCAP/MELD, with its self-cause ratio and trigger-localizing attention obtained directly from the forward pass, unifying how and why emotion changes in one mechanism.

REPRODUCIBILITY STATEMENT

We release the full implementation code under an MIT license, together with the pretrained multimodal features and the offline-extracted commonsense embeddings (subject to the IEMOCAP EULA and MELD CC BY 4.0 licenses). All hyperparameters are reported in Table 1 and the Appendix, all main results use a fixed seed (3407) with best-validation checkpoints, and stability is verified across three seeds. The complete prompt templates, JSON schemas, and extraction scripts for the offline commonsense and attribution-ratio annotations are included in the Appendix to enable independent reproduction.

ETHICS STATEMENT

This work uses public benchmark datasets (IEMOCAP and MELD) under their respective licenses; no new human-subject data is collected. The offline LLM annotation is used only to derive static reference features and a weak prior, without influencing individuals. We see no direct privacy, bias, or safety concern beyond those already documented for the underlying datasets and the pretrained encoders we reuse.

AI USE STATEMENT

We disclose the following use of generative AI tools (large language models).

- **Polishing writing.** We used an LLM to polish and improve the clarity, grammar, and readability of the manuscript.
- **Retrieval and discovery.** We used an LLM to assist literature search and the discovery of related work.
- **Drafting sections.** We used an LLM to draft initial versions of sections of the paper, which we subsequently reviewed, revised, and verified.
- **Synthetic datasets.** We used an LLM to generate synthetic annotations—specifically offline commonsense features (entity- and scene-level cues) and self/other attribution reference ratios—used as static pre-extracted features and a weak prior.

All LLM-produced annotations were generated under a fixed temperature-0 protocol and a strict JSON schema, spot-checked by human reviewers, and released together with the code. No language model is called during training or inference. We reviewed all AI-assisted work and take full responsibility for the final content of this work, including all text, claims, and artifacts.

REFERENCES

Daryl J. Bem. Self-perception theory. *Advances in Experimental Social Psychology*, 6:1–62, 1972.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations (ICLR)*, 2022.

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. *Language resources and evaluation*, 42:335–359, 2008.

Emily A Butler. Temporal interpersonal emotion systems: The "ties" that form relationships. *Personality and Social Psychology Review*, 15(4):367–393, 2011.

Yijing Dai, Jinxing Li, Yingjian Li, and Guangming Lu. Multi-modal graph context extraction and consensus-aware learning for emotion recognition in conversation. *Knowledge-Based Systems*, 298:111954, 2024.

Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462, 2010.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 154–164, 2019.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Findings of the association for computational linguistics: EMNLP 2020*, pp. 2470–2481, 2020.

James J. Gross. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299, 1998.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2122–2132, 2018.

Wen-Jue He, Xiaofeng Zhu, and Zheng Zhang. Cross-modal prompting for balanced incomplete multi-modal emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17463–17471, 2026.

Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7037–7041. IEEE, 2022.

Guimin Hu, Zhaohui Zhu, Daniel Hershcovich, Liming Hu, Hasti Seifi, and Jiayuan Xie. Unimeec: Towards unified multimodal emotion recognition and emotion cause. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5248–5261, 2024.

Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. Mmgcn: Multimodal fusion via deep graph convolution network for emotion recognition in conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5666–5675, 2021.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Junteng Jia and Austin R. Benson. Neural jump stochastic differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Ran Jing, Geng Tu, Yice Zhang, and Ruifeng Xu. Causal-erc: A multimodal framework with causal prompting for emotion recognition in conversations with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31383–31391, 2026.

Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 4148–4164, 2022.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. Emotional inertia and psychological maladjustment. *Psychological Science*, 21(7):984–991, 2010.

Jiang Li, Xiaoping Wang, Yingjian Liu, and Zhigang Zeng. Cfn-esa: A cross-modal fusion network with emotion-shift awareness for dialogue emotion recognition. *IEEE Transactions on Affective Computing*, 15(4):1919–1933, 2024.

Jiang Li, Xiaoping Wang, and Zhigang Zeng. Tracing intricate cues in dialogue: Joint graph structure and sentiment dynamics for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Xinran Li, Yu Liu, Jiaqi Qiao, and Xiujuan Xu. Do llms feel? teaching emotion recognition with prompts, retrieval, and curriculum learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31778–31786, 2026.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6818–6825, 2019.

Minh-Tien Nguyen, Huu-Loi Le, Manh-Cuong Phan, and Hajime Hotta. Cmtd: Cognitive modeling with traits and distortions for multimodal emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 839–854, 2026.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 873–883, 2017.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332, 2021.

Minjie Ren, Xiangdong Huang, Jing Liu, Zan Gao, Yuting Su, and An-An Liu. Ka-min: Knowledge-aware multimodal interaction network for emotion recognition in conversation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

Eunseon Seong, Harim Lee, Dahye Kim, Changhyun Kim, and Dong-Kyu Chae. Emotion-wheel-guided audio-referred text representation for multimodal emotion recognition in conversation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40392–40403, 2026.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1551–1560, 2021.

Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. Dynamic graph neural ordinary differential equation network for multi-modal emotion recognition in conversation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 256–268, 2025.

Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. Dual-constrained dynamical neural odes for ambiguity-aware continuous emotion prediction. *arXiv preprint arXiv:2407.21344*, 2024.

Tao Zhang and Zhenhua Tan. Ecerc: evidence-cause attention network for multi-modal emotion recognition in conversation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2064–2077, 2025.