Aligning Recency Across Modules for Multimodal Emotion-Cause Pair Extraction

Anonymous ACL submission

Abstract

Multimodal emotion-cause pair extraction (MECPE) is a structured link prediction problem that identifies emotion-cause utterance pairs under temporal precedence. While temporal proximity is a strong cue, modular MECPE architectures that mix sequential aggregation and speaker-interaction modules can encode inconsistent recency profiles across modules, destabilizing pair scoring. We propose **ATDG** (Adaptive Temporal Decay Generator), a low-capacity generator that maps *label-free* dialogue pace statistics to a dialogue-level time scale, and **DP** (Dual-Path Temporal Injection), which injects this shared scale into (*i*) **KS** (Kernel Smoothing), a kernel-smoothed sequential path that anchors pair scoring, and (*ii*) **SG** (Speaker Graph), a temporally decayed speaker-interaction graph path used only for emotion/cause prediction. Sharing a single timescale enforces cross-module temporal coherence without increasing model capacity. To protect the structured pair scorer under multi-task training, we adopt a pair-preserving two-stage schedule: Stage A learns the pair pathway under consistent temporal priors, and Stage B optionally refines the emotion/cause heads with the pair pathway frozen. Experiments on the ECF benchmark show consistent gains in pair extraction (up to 57.92 Pair F1) and robustness to evaluation-time perturbations of the guiding statistics. Code will be released publicly.

1 Introduction

Many neural systems combine heterogeneous reasoning modules, such as sequential context aggregation and graph-based interaction modeling (Gu et al., 2023; Li et al., 2024). Modular design enables flexibility and specialization, but it can also introduce inconsistent inductive biases across modules (Lippl and Lindsey, 2024; Zhang et al., 2024). A common source of inconsistency lies in how modules encode temporal recency (Chi et al., 2023; Hsieh et al., 2024). Sequential aggregators often smooth over long-range histories, whereas interaction graphs typically emphasize local, windowed context (Saha et al., 2023; Yao et al., 2023). When these modules are trained jointly, incompatible temporal assumptions can lead to unstable gradients and degraded performance on tasks that require coherent temporal reasoning (Lippl and Lindsey, 2024).

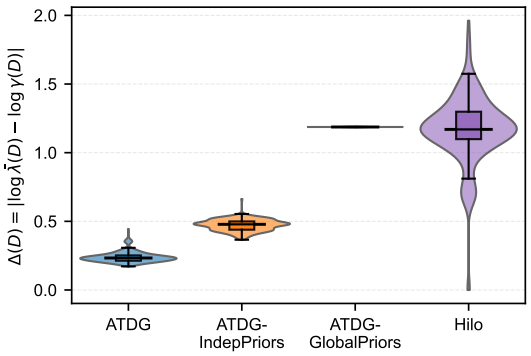

Figure 1: Per-dialogue cross-module temporal *misalignment* between the sequential aggregation and speaker-interaction components in MECPE models on the ECF test set (lower is better). Existing modular baselines (HiLo) exhibit higher median misalignment and larger dispersion, suggesting less coordinated effective temporal scales across components. In contrast, our method (ATDG), which explicitly shares a dialogue-level timescale, substantially reduces the magnitude and variance of this misalignment.

Understanding *why* an emotion arises in a conversation is central to explainable conversational systems (Yu et al., 2025a; Liu et al., 2025; Yu et al., 2025b). Emotion-cause pair extraction (ECPE) links an emotional utterance to its triggering antecedent utterances (Xia and Ding, 2019; Hu et al., 2024a). This task is further challenged in multi-speaker, multimodal dialogues (MECPE), where causal evidence may appear asynchronously across text, audio, and visual streams (Wang et al., 2023a; Hu et al., 2024b; Wang et al., 2024b).

A recurring challenge in modular MECPE systems is cross-module temporal misalignment. Dif-

ferent components can encode incompatible recency biases, yielding mismatched *temporal horizons*. Sequential aggregators smooth over long-range history, whereas speaker-interaction graphs focus on local exchanges (Smith et al., 2023; Qamar et al., 2023). When trained jointly, such mismatched temporal priors can increase gradient conflict and make pair ranking less stable, especially for non-local emotion-cause links.

To quantify this *misalignment*, we conduct a proxy cross-module temporal misalignment diagnostic, comparing within each dialogue the temporal scales of the sequential aggregation and interaction components (see Figure 1). This diagnostic suggests that representative modular MECPE baselines exhibit higher median misalignment and larger dispersion, whereas aligned temporal priors yield more coherent time-scale behavior.

To address this issue, our idea is to learn a dialogue-level time scale and share it across modules. We propose Adaptive Temporal Decay Generator (ATDG), which maps label-free dialogue-pace statistics to this latent scale and outputs positive decay rates for both sequential aggregation and speaker-graph message passing. We then inject the shared time scale through a Dual-Path temporal injection (DP): kernel smoothing performs order-respecting full-history aggregation as the sole input to the pair scorer, while the speaker graph refines utterance representations for utterance-level emotion/cause supervision without feeding the pair scorer. To further prevent auxiliary objectives from perturbing link ranking, we adopt a pair-preserving two-stage schedule that learns pair extraction first and then optionally refines utterance-level heads while keeping the pair pathway fixed.

Our contributions are three-fold. **First**, we introduce a dialogue-paced temporal prior that generates a shared time scale from label-free dialogue-pace statistics. **Second**, we propose a dual-path temporal injection that aligns the effective horizons of sequential aggregation and speaker interaction while keeping pair decoding clean. **Third**, we provide mechanism-level evidence and extensive evaluations on the ECF benchmark, including ablations and robustness analyses that validate the role of temporal alignment.

2 Related Work

Emotion-Cause Pair Extraction. ECPE formulates emotion understanding as structured link prediction under temporal precedence (Wang et al., 2024a, 2025). Recent work extends ECPE to multi-speaker, multimodal settings, establishing MECPE benchmarks such as ECF built on MELD (Poria et al., 2019; Wang et al., 2024b). Representative MECPE systems combine sequential context aggregation, speaker-aware interaction, and multimodal fusion for pair extraction (Li et al., 2025; Tu et al., 2026). While effective, these models treat temporal bias as a module-specific design choice, with each component encoding its own recency assumption. Prior work on multi-task MECPE reports negative transfer between utterance-level objectives and pair extraction (Su et al., 2024), typically addressed via loss balancing or architectural isolation. In contrast, we identify cross-module temporal mismatch as a root cause of such interference and address it by enforcing a shared, dialogue-paced temporal prior with a pair-preserving optimization protocol.

Temporal Biases in Conversational Reasoning. Temporal proximity is a core cue in ECPE and conversational modeling, commonly implemented via fixed windows, distance-aware priors, or position-based constraints (Zhou et al., 2024; Zhang et al., 2022). Broader long-context modeling explores learnable recency and memorization mechanisms to balance local salience and global context (Hou et al., 2023; Gu and Dao, 2023). However, existing methods introduce temporal bias within individual modules (e.g., attention, recurrence, or graph edges) without considering their interaction when composed (Wu et al., 2025b; Tu et al., 2024a). Consequently, different components may assume incompatible effective time scales, an issue particularly pronounced in modular MECPE architectures (Luo et al., 2024). Our approach departs from prior work by learning a low-capacity, dialogue-conditioned time-scale prior and consistently injecting it across heterogeneous mechanisms, aligning recency assumptions without increasing model expressiveness.

Speaker Interaction for Emotion Modeling. Speaker-aware modeling is central to conversational emotion analysis (Majumder et al., 2019; Song et al., 2023; Tu et al., 2024b; Wu et al., 2025a). Prior studies employ recurrent models, graph-based interaction networks, and multimodal fusion to capture cross-speaker influence (Hazarika et al., 2018; Shi and Huang, 2023; Yang et al., 2024). Graph-based speaker interaction has proven effective for MECPE (Tu et al., 2026), with most work focusing

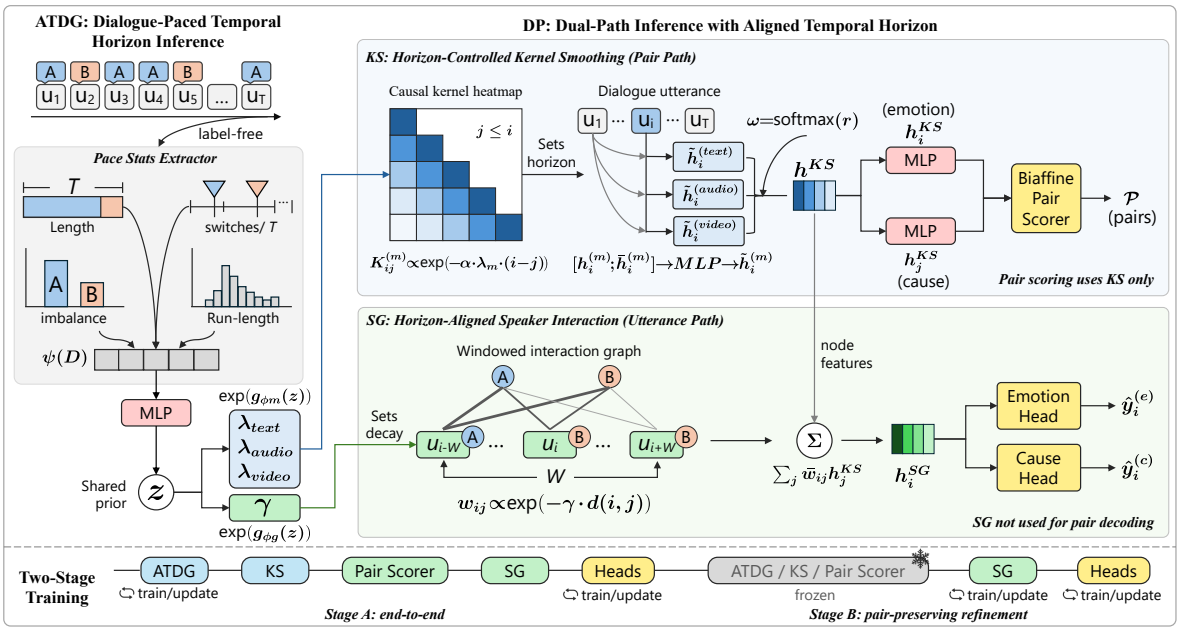

Figure 2: Overview of the proposed ATDG+DP framework. ATDG infers a dialogue-paced temporal horizon from label-free dialogue statistics. DP injects this shared horizon into kernel smoothing (KS) for pair scoring and a speaker graph (SG) for utterance-level prediction, with KS used exclusively for pair decoding. Training follows a pair-preserving two-stage schedule.

on interaction operators or fusion designs (Li et al., 2023; Yun et al., 2024). In contrast, we do not propose a speaker graph. Instead, we study how speaker interaction and sequential aggregation can be temporally aligned through a shared dialogue-paced prior, and how restricting speaker interaction to auxiliary heads enables stable optimization without contaminating structured pair decoding.

3 Method

3.1 Task Definition

We consider a dialogue D with T utterances. Each utterance t has multimodal inputs (x_t, a_t, v_t) (text, audio, video) and a speaker ID s_t. The goal is to jointly predict: (i) an emotion label $y_t^{(e)}$ for each utterance t, (ii) a binary cause indicator $y_t^{(c)} \in \{0, 1\}$ for each utterance t, and (iii) a set of utterance-indexed emotion-cause pairs $\mathcal{P} = \{(i, j)\}$ where utterance j triggers the emotion expressed at utterance i. Here $y_t^{(c)}$=1 indicates that utterance t serves as a cause for at least one emotion in the dialogue (i.e., it appears as the cause in some $(i, t) \in \mathcal{P}$). We assume a temporal precedence constraint ($j \leq i$) and enforce it with a lower-triangular mask during pair scoring. This formulation makes temporal distance Δ=$i-j$ a central structural variable, motivating consistent recency modeling across heterogeneous modules.

3.2 Inference Overview: From Temporal Mismatch to Aligned Reasoning Paths

Figure 1 shows that in modular MECPE models, the sequential aggregation path and the speaker-interaction path can operate under substantially different effective temporal scales within the same dialogue. Such cross-module mismatch is undesirable for structured emotion-cause reasoning: while pair decoding relies on coherent temporal ordering, inconsistent recency biases across paths can introduce unstable gradients and distort shared representations during joint training.

Our method addresses this issue by explicitly aligning temporal horizons across components. As shown in Figure 2, we first infer a dialogue-paced temporal scale that reflects how far relevant context should extend in the conversation. This single time scale is then applied consistently to both (i) the global sequential aggregation used for pair reasoning and (ii) the local speaker-interaction mechanism used for auxiliary refinement, while keeping the structured pair pathway clean.

Two paths, two roles. Given this aligned temporal horizon, our model separates inference into two complementary paths with distinct responsibilities. Kernel Smoothing (KS) produces $\mathbf{h}_t^{\mathrm{KS}}$ and provides the *only* input to the pair scorer. Speaker Graph (SG) takes $\mathbf{h}_t^{\mathrm{KS}}$ as node features and performs local cross-speaker interaction to pro-

duce $\mathbf{h}_t^{\mathrm{SG}}$ for utterance-level emotion detection and cause detection. SG is not used for pair scoring, whereby restricting it to auxiliary heads is crucial for our pair-preserving two-stage optimization. Consequently, any Pair F1 gains arise from KS and from improved training-time stability of $\mathbf{h}^{\mathrm{KS}}$ under auxiliary supervision rather than from SG directly participating in pair decoding.

3.3 Step I: Inferring a Dialogue-Paced Temporal Horizon

Given a dialogue D, our goal is to infer a single dialogue-level temporal horizon that reflects how far relevant causal evidence typically extends within this conversation. We describe ATDG, which derives a low-capacity, label-free dialogue-pace descriptor $\psi(D)$ as a low-dimensional vector summarizing surface-level dynamics without using semantic content. Concretely, $\psi(D)$ aggregates statistics such as: (i) dialogue length and speaker count, (ii) speaker-switching frequency, (iii) speaker participation imbalance, and (iv) run-length statistics of same-speaker turns. We then map $\psi(D)$ to a latent time-scale factor z via a lightweight MLP, and generate positive decay rates using a softplus parameterization:

$$\begin{aligned} z &= f_\theta(\psi(D)), \\ \lambda_m &= \mathrm{softplus}(g_{\phi_m}(z)), \\ \gamma &= \mathrm{softplus}(g_{\phi_g}(z)). \end{aligned} \tag{1}$$

The complete definition of $\psi(D)$ and all statistics are provided in Appendix A. ATDG operates on label-free dialogue statistics and does not access emotion, cause, or pair annotations. Although ATDG parameters are learned during training, the dialogue-level temporal horizon is inferred on-the-fly at inference time from label-free statistics.

3.4 Step II: Applying the Shared Temporal Horizon across Reasoning Paths

3.4.1 Pair-Level Sequential Reasoning with Horizon-Controlled Context

Sequential encoding. For each modality m, we first obtain utterance-level embeddings, and then apply a dialogue-level BiLSTM to model the utterance sequence. $\mathbf{H}^{(m)} = [\mathbf{h}_1^{(m)}, \cdots, \mathbf{h}_T^{(m)}] \in \mathbb{R}^{T \times d_h}$. MECPE is evaluated offline with full dialogue context. We enforce temporal precedence only in pair scoring (Section 3.1) and in the explicit distance-weighting terms below.

ATDG-controlled order-respecting decay kernel. We define an exponential decay kernel over temporal distance:

$$K_{ij}^{(m)} = \frac{\exp\big(-\alpha_{\mathrm{ker}}\,\lambda_m \cdot (i-j)\big) \cdot \mathbb{I}[j \le i]}{\sum_{k=1}^{T} \exp\big(-\alpha_{\mathrm{ker}}\,\lambda_m \cdot (i-k)\big) \cdot \mathbb{I}[k \le i]}, \tag{2}$$

where λ_m is generated by ATDG for modality m and α_{ker} is a global temperature shared across all settings. Here, α_{ker} controls kernel sharpness globally, while λ_m determines the dialogue-specific effective temporal horizon.

Kernel smoothing and fusion. We aggregate historical context as $\bar{\mathbf{h}}_i^{(m)} = \sum_{j=1}^{T} K_{ij}^{(m)}\,\mathbf{h}_j^{(m)}$, and fuse the base and smoothed states:

$$\tilde{\mathbf{h}}_i^{(m)} = \mathrm{MLP}\left([\mathbf{h}_i^{(m)}; \bar{\mathbf{h}}_i^{(m)}]\right). \tag{3}$$

We then combine modalities with a convex mixture $\boldsymbol{\omega} = \mathrm{softmax}(\mathbf{r})$, where $\mathbf{r} \in \mathbb{R}^M$ are mixture logits as $\mathbf{h}_i^{\mathrm{KS}} = \sum_{m=1}^{M} \omega_m \tilde{\mathbf{h}}_i^{(m)}$. KS anchors pair extraction and provides temporally smoothed representations controlled by the dialogue-paced prior.

3.4.2 Utterance-Level Refinement with Horizon-Aligned Speaker Interaction

We build a local speaker interaction graph within a window of size W. For any edge from utterance j to i (typically cross-speaker and within-window), we assign a time-decayed weight:

$$w_{ij} \propto \exp(-\gamma\, d(i,j)), \tag{4}$$

where $d(i,j)$ is the utterance distance and γ is generated by ATDG. We perform a lightweight, normalized message passing to obtain refined node representations $\tilde{h}_i$ from neighbors $\mathcal{N}(i)$:

$$\tilde{h}_i = \sum_{j \in \mathcal{N}(i)} \bar{w}_{ij}\, h_j, \tag{5}$$

where $\bar{w}_{ij}$ denotes normalized weights. We use $\tilde{h}_i$ only for utterance-level heads, while pair scoring relies on the KS pathway (Section 3.4.1). Please refer Appendix B for full SG weighting terms and update details.

3.5 Decoding Structured Emotion-Cause Links

Utterance-level emotion and cause. Given SG-refined representations $\mathbf{h}_{b,i}^{\mathrm{SG}}$, we compute logits for emotion and cause as

$$\begin{aligned} \hat{\mathbf{y}}_{b,i}^{(e)} &= \mathrm{MLP}_e(\mathbf{h}_{b,i}^{\mathrm{SG}}) \in \mathbb{R}^C, \\ \hat{\mathbf{y}}_{b,i}^{(c)} &= \mathrm{MLP}_c(\mathbf{h}_{b,i}^{\mathrm{SG}}) \in \mathbb{R}^2, \end{aligned} \tag{6}$$

where C is the number of emotion classes.

Pair scoring. We score each candidate pair (i, j) using only KS representations:

$$\mathbf{s}_{b,i,j} = \mathrm{Biaff}\big(\mathrm{MLP}_h(\mathbf{h}^{\mathrm{KS}}_{b,i}),\ \mathrm{MLP}_t(\mathbf{h}^{\mathrm{KS}}_{b,j})\big) \in \mathbb{R}^2, \tag{7}$$

where $\mathbf{s}_{b,i,j}$ are logits for $\{0, 1\}$ (non-pair vs. pair). SG is used only for utterance-level heads.

Masks. Dialogues are padded in a mini-batch. Let $m_{b,i} = \mathbb{I}[i \le L_b]$ be the utterance mask for dialogue b with length L_b. We define a valid-pair mask enforcing padding and temporal precedence:

$$M_{b,i,j} = \mathbb{I}[i \le L_b] \cdot \mathbb{I}[j \le L_b] \cdot \mathbb{I}[j \le i]. \tag{8}$$

We apply $M_{b,i,j}$ in both pair training and decoding.

3.6 Learning Objective with Causal Constraints

We train the model with a multi-task objective combining (i) emotion-cause pair classification and (ii) utterance-level emotion/cause prediction. For any dialogue, the pair scorer considers only forward links via a causal mask ($j \le i$), ensuring causes precede (or coincide with) the emotional response.

Let $\mathcal{L}_{\text{pair}}$ as the masked pairwise classification loss over all valid utterance pairs, and $\mathcal{L}_{\text{emo}}$ and $\mathcal{L}_{\text{cau}}$ as the utterance-level emotion and cause losses. We use standard (optionally class-weighted) cross-entropy for each term. The overall objective:

$$\mathcal{L} \;=\; \mathcal{L}_{\text{pair}} \;+\; \alpha\,\mathcal{L}_{\text{emo}} \;+\; \beta\,\mathcal{L}_{\text{cau}}, \tag{9}$$

where α and β balance utterance-level supervision against pair extraction. Full loss definitions, weighting details, and the explicit masking operator are provided in Appendix C.

3.7 Inference-Time Link Selection

At inference time, we compute $p_{b,i,j} = \mathrm{softmax}(\mathbf{s}_{b,i,j})[1]$ (masked to $j \le i$) and predict (i, j) if $p_{b,i,j} > \tau$, with τ selected on validation set.

3.8 Step III: Stabilizing Pair Inference under Auxiliary Supervision

We adopt a two-stage schedule to preserve pair extraction while benefiting from auxiliary refinement. **Stage A:** We optimize the full objective in Eq. (9) end-to-end. **Stage B:** We freeze the KS pathway and the pair scorer, and continue training only the SG pathway and utterance-level heads. This prevents auxiliary supervision from perturbing the learned pair-ranking function.

4 Experiments

4.1 Setup and Baselines

We evaluate **ECF** (Emotion-Cause-in-Friends) dataset (Wang et al., 2023a). Split statistics and label distributions are summarized in Appendix D. Following prior work, we adopt the offline evaluation protocol and report Precision/Recall/F1 for (i) emotion detection (macro-averaged), (ii) cause detection (binary positive class), and (iii) emotion-cause pair extraction (pair-level). Checkpoints are selected based on the best validation Pair F1.

We compare against heuristic position-statistics methods and a two-stage pipeline (Wang et al., 2023a), the hierarchical HiLo (Li et al., 2025), and prompt-only LLM baselines. Full hyperparameter and baseline details are in Appendix E.

4.2 Main Results

Overall performance on ECF. Table 1 reports results under the standard ECF protocol. Across all settings, ATDG+DP achieves the best Pair F1, indicating a dialogue-paced time-scale prior improves structured emotion-cause linking when used to parameterize temporal decay. Importantly, pair decoding uses only KS, while SG contributes indirectly by refining utterance-level heads and shaping representations during joint training in Stage A. In the full multimodal setting (Audio+Video), our model achieves **57.92** Pair F1, outperforming the strongest published baseline (HiLo, 55.45) by **+2.47** absolute F1. The gain is primarily driven by higher Pair Recall (65.12 vs. 59.69), consistent with improved recovery beyond immediate adjacency (see Section 4.6) without sacrificing precision. Pair AUPRC also increases (see Table 2), while precision remains stable and recall improves at longer distances.

Auxiliary tasks and multi-task stability. Although our design targets structured pair extraction, ATDG+DP remains competitive on utterance-level emotion and cause detection. Table 2 shows that **Stage A alone already achieves the best Pair F1**, indicating that pair structure is learned without reliance on auxiliary refinement. Stage B is therefore an optional, pair-preserving refinement step that improves emotion and cause predictions while leaving Pair F1 unchanged (see also Appendix F.1).

Effect of training schedules. To quantify the impact of joint optimization, we compare different training schedules in Table 2. Single-stage joint

Table 1: Test-set results across ECF subtasks. Baseline results are obtained from their original publications. For our **ATDG+DP** method, we report the mean and standard deviation $(\pm\text{std})$ across 3 random seeds. Best and second-best results are indicated in **bold** and underlined, respectively.

Methods		Emotion Detection			Cause Detection			Pair Extraction		
		P	R	F1	P	R	F1	P	R	F1
Heuristic Approach	$E_{\text{pred.}} + C_{\text{Bern.}}$	73.62	79.68	76.48	54.91	50.28	52.44	36.99	26.77	31.01
	$E_{\text{pred.}} + C_{\text{Multi.}}$	73.62	79.68	76.48	54.88	50.22	52.39	36.94	26.71	30.96
MECPE -2steps	Text-only	77.17	81.36	79.10	67.47	73.19	70.13	57.64	48.72	52.71
	+ Audio	76.91	81.68	79.17	67.25	73.91	70.27	57.13	50.34	53.48
	+ Video	77.10	81.18	79.03	68.10	72.51	70.18	**57.77**	49.53	53.21
	+ Audio + Video	77.45	81.10	79.16	68.42	72.43	70.27	57.47	49.81	53.20
HiLo	Text-only	74.53	81.80	77.79	65.26	80.23	71.97	51.91	54.33	53.09
	+ Audio	78.31	77.71	78.01	66.32	79.17	72.18	51.50	56.65	53.95
	+ Video	76.03	80.14	78.03	61.98	**85.89**	72.01	52.41	50.60	51.64
	+ Audio + Video	75.46	**83.66**	79.35	65.72	80.31	72.28	51.78	59.69	55.45
LLMs	LLaMA-3-8B	54.28	61.73	57.76	37.42	26.18	30.81	7.16	49.23	12.51
	Qwen-3-8B	59.14	66.85	62.76	41.26	30.47	35.06	9.73	51.68	16.38
ATDG+DP **(Ours)**	Text-only	$78.58_{\pm 0.76}$	$79.52_{\pm 1.00}$	$79.04_{\pm 0.12}$	$68.30_{\pm 0.95}$	$78.96_{\pm 1.84}$	$73.23_{\pm 0.24}$	$53.12_{\pm 0.89}$	$61.22_{\pm 0.56}$	$56.88_{\pm 0.50}$
	+ Audio	$79.11_{\pm 0.69}$	$79.70_{\pm 1.30}$	$\mathbf{79.40}_{\pm 0.42}$	$67.78_{\pm 0.24}$	$79.57_{\pm 1.69}$	$73.19_{\pm 0.59}$	$52.40_{\pm 2.86}$	$62.45_{\pm 4.29}$	$56.86_{\pm 1.05}$
	+ Video	$79.04_{\pm 0.26}$	$78.96_{\pm 0.25}$	$79.00_{\pm 0.23}$	$68.81_{\pm 0.42}$	$77.26_{\pm 2.29}$	$72.77_{\pm 0.80}$	$52.85_{\pm 2.35}$	$62.19_{\pm 2.59}$	$57.07_{\pm 0.74}$
	+ Audio + Video	$\mathbf{79.24}_{\pm 0.76}$	$79.45_{\pm 1.36}$	$79.33_{\pm 0.30}$	$\mathbf{69.12}_{\pm 0.65}$	$77.98_{\pm 2.47}$	$\mathbf{73.26}_{\pm 0.80}$	$52.16_{\pm 0.68}$	$\mathbf{65.12}_{\pm 1.71}$	$\mathbf{57.92}_{\pm 0.52}$

Table 2: Effect of training schedules on structured pair extraction (mean±std over 3 seeds), showing that single-stage joint training degrades Pair F1 while the pair-preserving schedule maintains pair performance.

Setting	Pair F1	Pair AUPRC	Emo. F1	Cause F1
ATDG+DP	57.92±0.52	55.58±0.57	79.33±0.30	73.26±0.80
-Stage A only	57.92±0.52	55.58±0.57	78.52±0.84	72.34±0.46
ATDG+DP-Joint	56.55±0.74	55.24±0.56	79.06±0.63	72.51±0.81

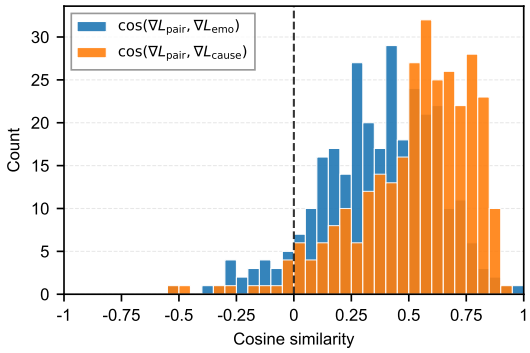

Figure 3: Gradient cosine similarity distributions under single-stage joint training (ATDG+DP-Joint), indicating increased gradient conflict between the pair objective and auxiliary objectives.

training (*ATDG+DP-Joint*) reduces Pair F1 by 1.37 pp relative to the pair-preserving schedule, despite achieving comparable utterance-level performance. In contrast, the two-stage schedule preserves Pair F1, while allowing Stage B to refine auxiliary heads without perturbing the pair pathway. These results demonstrate that optimization coupling has a measurable negative effect on structured pair learning, motivating explicit decoupling.

4.3 Diagnosing Optimization Conflict in Joint Training

While multi-task supervision improves utterance-level predictions, it may interfere with structured pair learning when all objectives are optimized jointly. To diagnose this effect, we analyze gradient interference under single-stage training.

For a training mini-batch, let $g_{\text{pair}} = \nabla_\theta \mathcal{L}_{\text{pair}}$, $g_{\text{e}} = \nabla_\theta \mathcal{L}_{\text{e}}$, and $g_{\text{c}} = \nabla_\theta \mathcal{L}_{\text{c}}$, where θ denotes parameters updated during joint training. We measure gradient alignment using cosine similarity $\cos(g_a, g_b) = \frac{g_a^\top g_b}{\|g_a\|_2 \, \|g_b\|_2}$, where negative values indicate conflicting update directions.

Figure 3 shows a non-trivial negative-cosine tail, with stronger conflict between the Pair and Emotion objectives than between Pair and Cause. This mechanism-level evidence aligns with the observed degradation under joint training and explains why decoupling auxiliary refinement preserves pair structure (see Appendix F.2 for details).

4.4 Core Ablations

Naming. Unless stated, all ablations keep DP unchanged. We denote prior-generation variants as **ATDG-GlobalPriors** (dialogue-agnostic) and **ATDG-IndepPriors** (separate latent factors for KS and SG). **-w/o KS** and **-w/o SG** disable DP time-scale injection on the KS and SG branches, respectively, while keeping the underlying KS/SG modules and training pipeline unchanged.

Hypotheses. Our ablation test three hypotheses: (H1) dialogue-conditioned priors outperform dialogue-agnostic priors; (H2) sharing one time

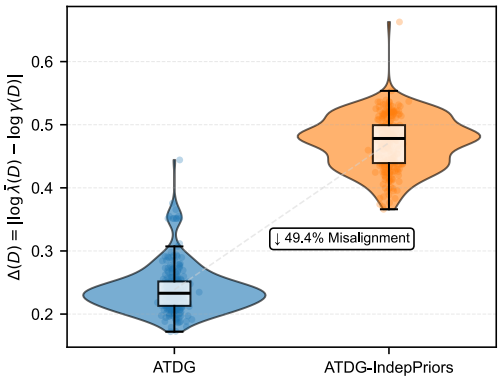

Figure 4: Per-dialogue cross-module temporal misalignment $\Delta(D)$ induced by shared vs. independent temporal priors on the ECF test set (lower is better).

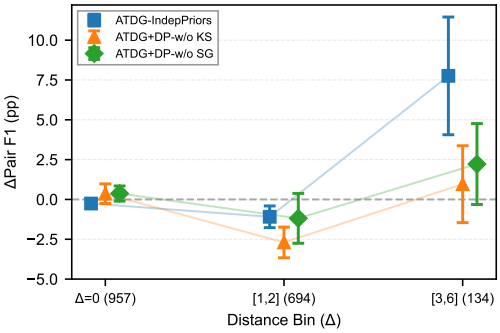

Figure 5: Pair F1 change (pp) relative to ATDG+DP across temporal distance bins on the ECF test set, highlighting behavior under sparse supervision (mean±std over 3 seeds).

scale across KS and SG is preferable to independent scales; and (H3) KS and SG provide complementary benefits.

Dialogue-adaptive vs. dialogue-agnostic priors (H1). As shown in Table 3, a comparison with the BASE model shows that ATDG+DP improves Pair F1 from **55.30** to **57.92** (+2.62), supporting the use of a learned time-scale prior as an inductive bias for the MECPE task. Dialogue-agnostic priors (ATDG-GlobalPriors) are beneficial but underperform full ATDG+DP in Pair F1, suggesting that dialogue-level adaptation captures meaningful variation in effective horizons. While ATDG-GlobalPriors slightly improves threshold-free Pair AUPRC, it underperforms in Pair F1 after validation-threshold selection, which is our primary metric.

Shared vs. independent priors (H2). Learning independent priors for KS and SG (ATDG-IndepPriors) yields a lower overall Pair F1 than sharing a single latent factor, supporting our central claim that consistent temporal scaling reduces mismatch across modules. Notably, while some

Table 3: Core ablations on ECF test set over 3 seeds, showing that dialogue-conditioned and shared temporal priors provide the largest gains for pair extraction.

Variant	Pair F1	Pair AUPRC	Emo. F1	Cause F1
ATDG+DP (full)	**57.92**$_{\pm 0.52}$	55.58$_{\pm 0.57}$	**79.33**$_{\pm 0.30}$	**73.26**$_{\pm 0.80}$
-w/o KS	56.79$_{\pm 0.24}$	52.92$_{\pm 2.18}$	78.89$_{\pm 0.44}$	73.17$_{\pm 0.32}$
-w/o SG	57.23$_{\pm 0.79}$	55.46$_{\pm 0.81}$	78.99$_{\pm 0.46}$	72.97$_{\pm 0.42}$
ATDG-IndepPriors	56.96$_{\pm 0.51}$	55.66$_{\pm 0.19}$	79.13$_{\pm 1.09}$	72.88$_{\pm 0.56}$
ATDG-GlobalPriors	56.78$_{\pm 0.05}$	**56.24**$_{\pm 0.10}$	79.10$_{\pm 0.24}$	72.45$_{\pm 0.23}$
BASE (no ATDG/DP)	55.30$_{\pm 0.16}$	53.20$_{\pm 0.68}$	78.69$_{\pm 0.20}$	72.43$_{\pm 0.55}$

long-tail bins may fluctuate (Section 4.6), shared priors provide more reliable overall performance.

Dual-path complementarity (H3). Removing either KS or SG degrades Pair F1. KS contributes most to pair extraction, consistent with its role as the primary sequential aggregator for link prediction, while SG provides complementary gains via speaker-aware refinement and KS stabilization, despite being excluded from decoding.

4.5 Diagnosing Cross-Module Temporal Alignment

Our core claim is that a shared, dialogue-paced time scale aligns the notion of recency between the sequential KS path and the speaker-graph SG path. As a direct mechanism-level diagnostic, for each dialogue D we measure the cross-module temporal misalignment $\Delta(D) = \left|\log \bar{\lambda}(D) - \log \gamma(D)\right|$, where $\bar{\lambda}(D)$ is the mean KS decay across modalities and $\gamma(D)$ is the SG decay (Appendix F.3). Lower $\Delta(D)$ means two paths yield more consistent effective horizons for the same dialogue.

Figure 4 shows that ATDG+DP achieves the lowest misalignment across dialogues. In contrast, ATDG-IndepPriors increases $\Delta(D)$ when KS and SG are driven by separate latent factors, and dialogue-agnostic priors (ATDG-GlobalPriors), as well as the HiLo baseline exhibit substantially larger mismatch. These results demonstrate that sharing a dialogue-conditioned time scale improves temporal coherence across heterogeneous modules.

4.6 Distance-Specific Behavior

Where do the gains come from? Figure 5 compares several ATDG/DP-related variants against ATDG+DP across temporal distance bins. Most gold pairs lie at short distances, where all variants perform similarly due to strong local regularities. Differences emerge at larger distances where supervision becomes sparse.

Stability in sparse regimes. At medium distances, removing KS yields the largest drop, consis-

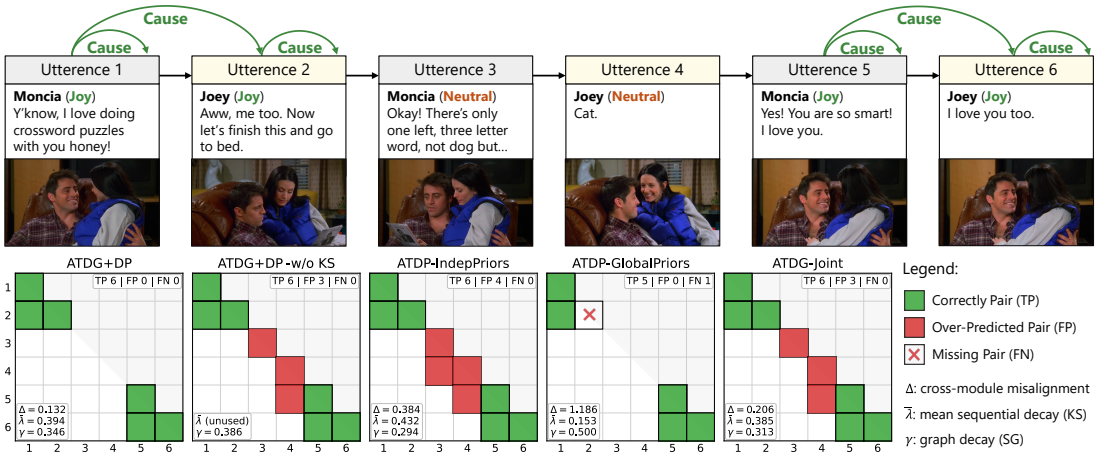

Figure 6: Pair-level case study on ECF illustrating how dialogue-adaptive temporal decay affects pair ranking. **Top:** a dialogue with gold emotion–cause links. **Bottom:** predicted pair matrices for different variants. Rows index emotion utterances i, columns index candidate causes j.

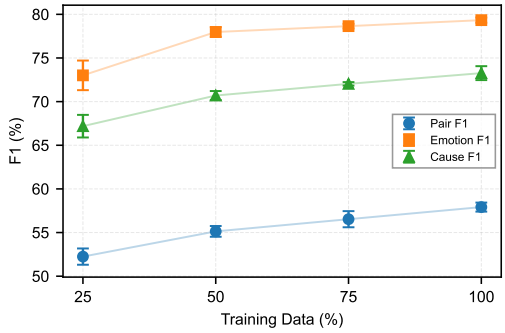

Figure 7: Data efficiency on the ECF test set (mean±std over 3 seeds). Larger gains at lower data fractions suggest the dialogue-paced prior improves data efficiency.

tent with KS functioning as a kernel smoother controlled by ATDG decay. In the long-tail bin, some variants can show higher mean performance but substantially larger variance. For example, ATDG-IndepPriors exhibits larger error bars, suggesting that decoupled time scales may amplify noise when evidence is weak. Overall, the distance analysis demonstrates that *shared* temporal priors are valuable for reliable behavior under sparse supervision.

4.7 Case Study

Figure 6 shows pair-level errors in one dialogue. ATDG+DP recovers all gold links, including the self-link $(2, 2)$ and the cross-utterance link $(6, 5)$. ATDG-GlobalPriors misses $(2, 2)$, showing that a dialogue-agnostic horizon can under-link even near the diagonal. Conversely, weakening KS horizon control (ATDG+DP-w/o KS and ATDG+DP-Joint) introduces spurious links such as $(5, 4)$ in the neutral segment. ATDG-IndepPriors exhibits similar false positives, consistent with a mismatch in recency between the KS and SG paths.

4.8 Data Efficiency

Figure 7 evaluates data efficiency by training with increasing fractions of the ECF training set. Pair F1 improves from **52.24** at 25% data to **57.92** at full data (Appendix F.7), with the steepest gains on pair extraction. This pattern is consistent with ATDG+DP providing a useful temporal inductive bias in low-data regimes while remaining compatible with additional supervision.

Robustness analyses. We further analyse eval-only sanity perturbations of dialogue statistics and missing-modality robustness (see Appendix F.5-F.6). Pair extraction degrades smoothly, indicating that the learned temporal priors function as stable inductive biases rather than brittle shortcuts, consistent with shared dialogue-level temporal alignment. See Appendix F.10 for real-time evaluation.

5 Conclusion

We investigate how temporal priors affect the extraction of multimodal emotion–cause pairs in conversations. This paper proposes ATDG to generate a dialogue-paced time scale from label-free dialogue statistics and DP to inject it consistently into a sequential kernel-smoothing path and a speaker-interaction graph. On ECF, this improves pair extraction up to 57.92 Pair F1, primarily via higher recall. A pair-preserving two-stage schedule further improves utterance-level prediction without degrading pair extraction. Future work includes evaluation on longer, diverse conversational benchmarks and enriching the pace descriptor with semantic cues such as topic shifts and discourse structure.

Limitations

We note several limitations and opportunities for extension. First, ATDG conditions on coarse, label-free pace statistics from speaker turns. This is a deliberate choice to keep the prior semantics-free and lightweight, but it may miss content-dependent signals such as topic shifts or implicit discourse relations that can modulate causal horizons. Incorporating richer yet non-leaky descriptors is a promising direction.

Second, SG models cross-speaker influence within a fixed local window W. This design emphasizes local interaction patterns, whereas longer-range speaker influence is handled implicitly through KS for pair scoring. Extending SG with adaptive or multi-scale neighborhoods could further improve robustness in longer dialogues.

Finally, we rely on pre-extracted acoustic and visual features rather than end-to-end multimodal encoders. We view encoder advances as complementary to our temporal-prior framework, and integrating stronger modality encoders is left to future work.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Ta-Chung Chi, Ting-Han Fan, Alexander Rudnicky, and Peter Ramadge. 2023. Dissecting transformer length extrapolation via the lens of receptive field analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13522–13537, Toronto, Canada. Association for Computational Linguistics.

Jesse Davis and Mark H. Goadrich. 2006. The relationship between precision-recall and ROC curves. In *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, volume 148 of *ACM International Conference Proceeding Series*, pages 233–240. ACM.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Jia-Chen Gu, Zhenhua Ling, Quan Liu, Cong Liu, and Guoping Hu. 2023. GIFT: Graph-induced fine-tuning for multi-party conversation understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11645–11658, Toronto, Canada. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. 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, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Guiyang Hou, Yongliang Shen, Wenqi Zhang, Wei Xue, and Weiming Lu. 2023. Enhancing emotion recognition in conversation via multi-view feature alignment and memorization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12651–12663, Singapore. Association for Computational Linguistics.

Cheng-Yu Hsieh, Yung-Sung Chuang, Chun-Liang Li, Zifeng Wang, Long Le, Abhishek Kumar, James Glass, Alexander Ratner, Chen-Yu Lee, Ranjay Krishna, and Tomas Pfister. 2024. Found in the middle: Calibrating positional attention bias improves long context utilization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14982–14995, Bangkok, Thailand. Association for Computational Linguistics.

Guimin Hu, Yi Zhao, and Guangming Lu. 2024a. Unifying emotion-oriented and cause-oriented predictions for emotion-cause pair extraction. *Neural Networks*, 178:106431.

Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. 2024b. UniMEEC: Towards unified multimodal emotion recognition and emotion cause. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5248–5261, Miami, Florida, USA. Association for Computational Linguistics.

Bobo Li, Hao Fei, Fei Li, Tat-Seng Chua, and Donghong Ji. 2025. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(11):307:1–307:19.

Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. Joyful: Joint modality fusion and graph contrastive learning for multimoda emotion recognition. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.

Jingyang Li, Shengli Song, Yixin Li, Hanxiao Zhang, and Guangneng Hu. 2024. Chatmdg: A discourse parsing graph fusion based approach for multi-party dialogue generation. *Inf. Fusion*, 110:102469.

Samuel Lippl and Jack W. Lindsey. 2024. Inductive biases of multi-task learning and finetuning: multiple regimes of feature reuse. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Hao Liu, Runguo Wei, Geng Tu, Jiali Lin, Dazhi Jiang, and Erik Cambria. 2025. Knowing what and why: Causal emotion entailment for emotion recognition in conversations. *Expert Systems With Applications*, 274:126924.

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

Meng Luo, Han Zhang, Shengqiong Wu, Bobo Li, Hong Han, and Hao Fei. 2024. NUS-emo at SemEval-2024 task 3: Instruction-tuning LLM for multimodal emotion-cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1589–1596, Mexico City, Mexico. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The*

Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pages 6818–6825. Proceedings of the AAAI Conference on Artificial Intelligence.

Daniel Mendoza, Francisco Romero, and Caroline Trippel. 2024. Model selection for latency-critical inference serving. In *Proceedings of the Nineteenth European Conference on Computer Systems, EuroSys 2024, Athens, Greece, April 22-25, 2024*, pages 1016–1038. ACM.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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

Ayesha Qamar, Adarsh Pyarelal, and Ruihong Huang. 2023. Who is speaking? speaker-aware multiparty dialogue act classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10122–10135, Singapore. Association for Computational Linguistics.

Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. 2023. Learning adaptive neighborhoods for graph neural networks. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 22484–22493. IEEE.

Tao Shi and Shao-Lun Huang. 2023. MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14752–14766, Toronto, Canada. Association for Computational Linguistics.

Jimmy T.H. Smith, Shalini De Mello, Jan Kautz, Scott W. Linderman, and Wonmin Byeon. 2023. Convolutional state space models for long-range spatiotemporal modeling. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Rui Song, Fausto Giunchiglia, Lida Shi, Qiang Shen, and Hao Xu. 2023. SUNET: speaker-utterance interaction graph neural network for emotion recognition in conversations. *Eng. Appl. Artif. Intell.*, 123(Part B):106315.

Xinxin Su, Zhen Huang, Yixin Su, Bayu Distiawan Trisedya, Yong Dou, and Yunxiang Zhao. 2024. Hierarchical shared encoder with task-specific transformer layer selection for emotion-cause pair extraction. *IEEE Transactions on Affective Computing*, 15(4):1934–1948.

Llama Team. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Geng Tu, Jun Wang, Zhenyu Li, Shiwei Chen, Bin Liang, Xi Zeng, Min Yang, and Ruifeng Xu. 2024a. Multiple knowledge-enhanced interactive graph network for multimodal conversational emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3861–3874, Miami, Florida, USA. Association for Computational Linguistics.

Geng Tu, Jun Wang, Li Yang, Bin Liang, Erik Cambria, Wenjie Li, and Ruifeng Xu. 2026. Multi-task mutual learning for multimodal emotion-cause pair extraction in conversations. *Information Fusion*, 127:103877.

Geng Tu, Tian Xie, Bin Liang, Hongpeng Wang, and Ruifeng Xu. 2024b. Adaptive graph learning for multimodal conversational emotion detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19089–19097. AAAI Press.

Botao Wang, Keke Tang, and Peican Zhu. 2024a. Enhancing emotion-cause pair extraction in conversations via center event detection and reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10773–10783, Miami, Florida, USA. Association for Computational Linguistics.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024b. SemEval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2039–2050,

Mexico City, Mexico. Association for Computational Linguistics.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023b. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3952–3963, Singapore. Association for Computational Linguistics.

Yuwei Wang, Yuling Li, Kui Yu, and Jing Yang. 2025. A semantic structure-based emotion-guided model for emotion-cause pair extraction. *Pattern Recognition*, 161:111296.

ChengYan Wu, Yiqiang Cai, Yang Liu, Pengxu Zhu, Yun Xue, Ziwei Gong, Julia Hirschberg, and Bolei Ma. 2025a. Multimodal emotion recognition in conversations: A survey of methods, trends, challenges and prospects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6257–6274, Suzhou, China. Association for Computational Linguistics.

Dan Wu, Xincheng Ju, Dong Zhang, Shoushan Li, Erik Cambria, and Guodong Zhou. 2025b. Emotion across modalities and cultures: Multilingual multimodal emotion-cause analysis with memory-inspired framework. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 5775–5783, New York, NY, USA. Association for Computing Machinery.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 40 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Zhenyu Yang, Xiaoyang Li, Yuhu Cheng, Tong Zhang, and Xuesong Wang. 2024. Emotion recognition in conversation based on a dynamic complementary graph convolutional network. *IEEE Transactions on Affective Computing*, 15(3):1567–1579.

Tianjun Yao, Yingxu Wang, Kun Zhang, and Shangsong Liang. 2023. Improving the expressiveness of k-hop message-passing gnns by injecting contextualized substructure information. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3070–3081, New York, NY, USA. Association for Computing Machinery.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2025a. Beyond verbal cues: Emotional contagion graph network for causal emotion entailment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1755–1767, Vienna, Austria. Association for Computational Linguistics.

Zhaoxin Yu, Xinglin Xiao, and Wenji Mao. 2025b. One unified model for diverse tasks: Emotion cause analysis via self-promote cognitive structure modeling. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10278–10293, Albuquerque, New Mexico. Association for Computational Linguistics.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. TelME: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 82–95, Mexico City, Mexico. Association for Computational Linguistics.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. TSAM: A two-stream attention model for causal emotion entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6762–6772, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Ruiyi Zhang, Xaq Pitkow, and Dora E. Angelaki. 2024. Inductive biases of neural network modularity in spatial navigation. *SCIENCE ADVANCES*, 10(29).

Zhihan Zhou, Xue Gu, Yujie Zhao, and Hao Xu. 2024. POP-CEE: Position-oriented prompt-tuning model for causal emotion entailment. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4199–4210, Bangkok, Thailand. Association for Computational Linguistics.

A Dialogue Pace Statistics

We provide the full definition of $\psi(D)$, including the exact normalization, entropy/imbalance measure, and run-length aggregation used in the main paper.

Label-free pace statistics. We summarize each dialogue D with a low-dimensional, label-free pace vector $\psi(D) \in \mathbb{R}^{d_s}$:

$$\psi(D) = \big[\log T,\ |S|,\ \mathrm{SR}(D),\ H(\mathbf{p}),\ \boldsymbol{\kappa}(D),\ \iota(\mathbf{p})\big], \tag{10}$$

where $d_s = 8$, T is the number of utterances, and $|S|$ is the number of unique speakers. $\mathrm{SR}(D) \in [0,1]$ is the fraction of adjacent speaker changes. Let $\mathbf{p} \in \mathbb{R}^{|S|}$ be the per-speaker utterance proportion (i.e., $p_k = \#\mathrm{utt}(s{=}k)/T$ and $\sum_k p_k = 1$) and $H(\mathbf{p}) = -\sum_k p_k \log p_k$ its entropy. Let $\boldsymbol{\ell}(D)$ be the run-length sequence of consecutive turns by the same speaker. We define $\boldsymbol{\kappa}(D) = \big[\mu_\ell,\ \ell_{\max},\ \sigma_\ell\big] \in \mathbb{R}^3$, with $\mu_\ell = \mathrm{mean}(\boldsymbol{\ell})$, $\ell_{\max} = \max(\boldsymbol{\ell})$, and $\sigma_\ell = \mathrm{std}(\boldsymbol{\ell})$. Finally, we measure speaker-imbalance as a scalar

$$\iota(\mathbf{p}) = 1 - \frac{H(\mathbf{p})}{\log |S|} \in [0,1], \tag{11}$$

where larger values indicate more uneven participation. These statistics are semantics-agnostic and encode conversational rhythm (turn-taking pace and participation inequality), providing a stable, low-capacity control signal.

Latent time-scale factor. ATDG maps $\psi(D)$ to a latent vector $\mathbf{z} \in \mathbb{R}^{d_z}$:

$$\mathbf{z} = f_\phi\big(\psi(D)\big), \tag{12}$$

where f_ϕ is a small MLP (low capacity by design) that does not predict links, but only parameterizes temporal decay.

One latent factor, two parameterizations. From the same $\mathbf{z}$, ATDG outputs (i) modality-specific sequential decay rates for KS and (ii) a graph-side decay rate for SG:

$$\boldsymbol{\lambda} = g_{\mathrm{seq}}(\mathbf{z}) \in \mathbb{R}^M; \quad \gamma = g_{\mathrm{graph}}(\mathbf{z}) \in \mathbb{R}. \tag{13}$$

where $M = 3$ modalities. We enforce positivity with $\mathrm{softplus}$:

$$\boldsymbol{\lambda} = \mathrm{softplus}(\mathbf{W}_{\mathrm{seq}}\mathbf{z} + \mathbf{b}_{\mathrm{seq}}), \tag{14}$$

$$\gamma = \mathrm{softplus}(\mathbf{w}_{\mathrm{graph}}^\top \mathbf{z} + b_{\mathrm{graph}}). \tag{15}$$

Sharing $\mathbf{z}$ enforces a coherent dialogue-level time scale across modules, while separate projections allow KS and SG to map the same latent horizon into their native decay parameterizations.

B Speaker Graph Details

We provide the full SG edge weighting function and the original message passing update used in our implementation.

Local speaker graph. For each utterance i, SG connects to a local set of preceding utterances within a window W:

$$\mathcal{N}(i) = \{j \mid 1 \le i - j \le W\}. \tag{16}$$

Here, W serves as a computational locality cap, while the learned decay (below) controls the relative influence profile within this cap. We intentionally keep SG local (window W) to model nearby cross-speaker influence, while global history is handled by KS through full-history kernel smoothing.

ATDG-controlled edge decay. For an edge $(j \to i)$ with temporal distance $d = i - j$, we apply an exponential decay $\delta(d) = \exp(-\gamma \cdot d)$, where γ is produced by ATDG from the same latent factor $\mathbf{z}$ used in KS.

Cross-speaker propagation for refinement. Using KS representations as node features, we compute a scalar interaction weight:

$$w_{ij} = \sigma\big(g([\mathbf{h}_j^{\mathrm{KS}}; \mathbf{h}_i^{\mathrm{KS}}])\big) \cdot \sigma(\rho_{s_i}) \cdot \delta(i-j) \cdot \mathbb{I}[s_i \neq s_j], \tag{17}$$

where $g(\cdot)$ is an MLP that outputs a scalar relatedness logit, and $\rho_{s_i} \in \mathbb{R}$ is a learned speaker susceptibility scalar for speaker s_i. We predict an influence vector from the source node $\mathbf{v}_j = \mathbf{W}_v \mathbf{h}_j^{\mathrm{KS}}$, and update representations via normalized message passing:

$$\mathbf{h}_i^{\mathrm{SG}} = \mathbf{h}_i^{\mathrm{KS}} + \eta \cdot \frac{\sum_{j \in \mathcal{N}(i)} w_{ij} \cdot \mathbf{v}_j}{\epsilon + \sum_{j \in \mathcal{N}(i)} w_{ij}}, \tag{18}$$

where η controls interaction strength and ϵ is a small constant for numerical stability. SG is used to refine utterance-level predictions while keeping the pair pathway stable (Section 3.8).

C Loss Details

Causal mask. We define a binary mask $M_{ij} = \mathbb{I}[j \le i]$ to restrict training and inference to forward links.

Table 4: Split-wise dataset statistics of ECF.

Metric	Train	Valid	Test
Conversations	1,001	112	261
Utterances	9,966	1,087	2,566
Emotion-Cause Pairs	7,055	866	1,873
Avg. Utterances/Conv.	9.96	9.71	9.83
Avg. Pairs/Conv.	7.05	7.73	7.18
Speakers	265	48	105
Avg. Utterance Length	11.19	11.33	11.48
Avg. Pair Distance	0.85	0.98	0.83

Pair loss. For each utterance pair (i, j), we compute logits s_{ij} and optimize a masked cross-entropy:

$$\mathcal{L}_{\text{pair}} = \sum_i \sum_j M_{ij} \cdot \text{CE}\big(y_{ij},\, s_{ij}\big), \qquad (19)$$

where CE may optionally use class weights to address imbalance.

Utterance losses. We apply cross-entropy for utterance-level emotion and cause prediction:

$$\mathcal{L}_{\text{emo}} = \sum_i \text{CE}(y_i^e, \hat{y}_i^e), \quad \mathcal{L}_{\text{cau}} = \sum_i \text{CE}(y_i^c, \hat{y}_i^c). \qquad (20)$$

D Dataset Statistics

Table 4 summarizes split-wise statistics of **ECF** (Emotion-Cause-in-Friends) (Wang et al., 2023a), a multimodal conversational emotion-cause dataset built on MELD (Poria et al., 2021). The ECF dataset is derived from publicly broadcast TV sitcoms. It is distributed under the GNU General Public License v3.0, and our use of the data is consistent with its intended research purpose.

Two properties are particularly relevant to our method. First, the average pair distance is below 1 utterance, confirming that temporal proximity is a strong cue on this benchmark. However, the existence of non-local pairs (and the long-tail bins in Appendix F.4) motivates learning a non-universal temporal scale rather than relying on a fixed window. Second, ECF exhibits structured sparsity at the pair level (pairs per conversation are modest relative to utterance counts), which makes calibrated link prediction important and partially explains why generation-style formulations for emotion-cause extraction (e.g., ECTEC as index generation) may over-predict links under sparse supervision (Wang et al., 2023b).

E Implementation Details

E.1 Evaluation Metrics

We follow the standard ECF evaluation protocol for emotion detection, cause detection, and pair extraction. Emotion detection is evaluated with Macro-F1 (F1) over seven classes to account for label imbalance. Cause detection is evaluated as binary classification and we report positive-class F1. Pair extraction evaluates whether predicted links match gold emotion-cause pairs under causal constraints ($j \leq i$). We report Pair F1, and additionally Pair AUPRC computed as Average Precision (AP) over all valid pairs using the pair scores (Davis and Goadrich, 2006).

Table 5: Key hyperparameters for reproducibility.

Category	Hyperparameter	Value
Backbone	Text Encoder	RoBERTa-base
	Audio Encoder	wav2vec2-base
	Video Encoder	VideoMAE-base
	Video Frames	16 (uniform)
	Multimodal Dim (d_a, d_v)	768
	Max Utterances	35
Architecture	Hidden Size d_{hid}	400
	Dropout	0.1
	Batch Size	32
	Loss weights (α, β)	1.0, 1.0
ATDG	Latent Dim d_z / Stats Dim	32 / 8
	KS: Kernel Temperature α_{ker}	1.0
	SG: Window W	6
	SG: Interaction Strength	0.02
Stage A	Text Encoder LR	1e-5
	Other LR	3e-5
	Weight Decay	1e-4
	Warmup Proportion	0.1
	Epochs / Patience	30 / 5
Stage B	Text Encoder LR	0 (frozen)
	Head / Other LR	3e-5 / 3e-4
	Epochs / Patience	10 / 3

E.2 Model Hyperparameters

We use RoBERTa-base (Liu et al., 2019) as the text encoder. For audio and video, we pre-extract utterance-level features offline using wav2vec 2.0-base (Baevski et al., 2020) and VideoMAE-base (Tong et al., 2022), respectively. We mean-pool encoder outputs to obtain 768-d vectors per utterance (16 uniformly sampled frames for video), which are then projected to the model hidden size.

E.3 Experimental Environment

All experiments were run on an AMD Ryzen 9 9950X CPU and an NVIDIA GeForce RTX 5090

Emotion–Cause Pair Extraction (Text-only)
You are given a dialogue consisting of multiple utterances. Each utterance has: (1) an index (starting from 0), (2) a speaker name/ID, and (3) the utterance text.
Your goal is to extract emotion–cause pairs from this dialogue.

1. **Emotion Label Prediction**
   Assign **exactly one** emotion label to **each** utterance from the following set:

   $$\{\texttt{neutral}, \texttt{surprise}, \texttt{anger}, \texttt{sadness}, \texttt{joy}, \texttt{disgust}, \texttt{fear}\}.$$

   Return the labels as a list in the same order as the utterances (length T).

2. **Cause Identification**
   Identify utterances that serve as **causes** for emotions expressed in other utterances. A cause is an utterance that provides evidence or an event/reason that triggers an emotion. A cause can be the same utterance as the emotion (*self-cause* is allowed). List the indices of all utterances that are causes (unique indices).

3. **Emotion–Cause Pairing**
   For each utterance i that expresses a **non-neutral** emotion, find its most plausible cause utterance(s) j. Output pairs as `[i, j]` where:

   - i is the index of the **emotion** utterance,
   - j is the index of the corresponding **cause** utterance,
   - the temporal precedence constraint must hold: $j \leq i$.

   If an emotion utterance has no clear cause, do not output a pair for it.

Output format (strict).
Return **only** a valid JSON object (no extra text) with the following keys:
- `"emotion"`: a list of T emotion strings (one per utterance),
- `"cause"`: a list of unique cause indices (integers),
- `"pair"`: a list of pairs `[i, j]` (integers) satisfying $j \leq i$.

Dialogue input format.
Each utterance is presented as: `[index] speaker: text`
Dialogue:

```
[0] SpeakerA: ...
[1] SpeakerB: ...
[2] SpeakerA: ...
...
```

Figure 8: Zero-shot prompt used for prompt-only LLM baselines in our text-only setting.

GPU with 32 GB memory. The software stack used PyTorch 2.8.0, CUDA 12.8, and Transformers 4.52.0. Unless otherwise stated, we fix random seeds to $\{42, 3456, 5678\}$ to measure variance under different initializations. We report the full software stack to facilitate reproducibility. The key hyperparameters for model architecture, and decoding are summarized in Table 5.

E.4 Baselines

Heuristic baselines. We include two position-statistics baselines from (Wang et al., 2023a), which combine a pretrained emotion classifier with probabilistic cause selection (Bernoulli vs. Multinomial).

Two-stage pipeline (MECPE-2steps). Following (Wang et al., 2023a), this pipeline first predicts utterance-level emotions and causes, and then scores candidate pairs with a bilinear matcher.

HiLo. We compare against HiLo (Li et al., 2025), a hierarchical model that performs dialogue-level reasoning on top of utterance encoders.

Prompt Template for LLM Baselines As introduced in Section 4.1, we include LLaMA-3-8B (Team, 2024) and Qwen-3-8B (Yang et al., 2025) as prompt-only LLM baselines for zero-shot Emotion–Cause Pair Extraction. Since these models operate on text, we evaluate them in a text-only setting (no audio/video inputs).

We use a structured zero-shot prompt that guides the model to: (i) assign an emotion label to each utterance, (ii) identify which utterances act as causes, and (iii) produce emotion–cause pairs while respecting temporal precedence ($j \leq i$). To support automatic scoring, we require the model to output a single JSON-formatted dictionary.

Figure 8 shows the exact prompt template used in our experiments. For each dialogue, we replace the placeholder with indexed utterances and

Table 6: Gradient conflict statistics during single-stage joint training (**ATDG+DP**). Values are mean±std over 3 seeds, computed from 100 training mini-batches per seed (diagnostic batch size 4).

Gradient pair	Mean cosine	Median cosine	% steps with $\cos < 0$
$\nabla\mathcal{L}_{\text{pair}}$ vs. $\nabla\mathcal{L}_{\text{e}}$	0.368 ± 0.016	0.385 ± 0.015	8.0 ± 4.1
$\nabla\mathcal{L}_{\text{pair}}$ vs. $\nabla\mathcal{L}_{\text{c}}$	0.530 ± 0.042	0.582 ± 0.024	3.7 ± 2.9

speaker names in the format `[index] speaker: text`.

F Additional Results

F.1 Two-Stage Training vs. Single-Stage

We ablate the training schedules described in Section 3.8. In addition to the full two-stage procedure, we report (i) *Stage-A-only* (early stop after Stage A), and (ii) *single-stage joint* training (**ATDG+DP**), which optimizes all objectives and all trainable parameters throughout.

Pair-preserving refinement vs. joint optimization. Table 2 shows that Stage A alone already matches the full two-stage model on Pair F1, indicating that the pair structure can be learned reliably in Stage A under the proposed temporal priors. Stage B then improves utterance-level metrics (Emotion/Cause F1) while keeping Pair F1 unchanged, consistent with a refinement step that does not perturb the learned pair scorer. In contrast, single-stage joint training yields a lower Pair F1, suggesting negative transfer from auxiliary objectives when all losses compete throughout optimization.

F.2 Extended Gradient Conflict Analysis

This section provides additional quantitative details for the gradient interference analysis reported in Section 4.3 (main paper). We report exact statistics and sampling protocols used to estimate gradient cosine distributions under single-stage joint training.

Table 6 shows a non-trivial negative-cosine tail, with stronger conflict between Pair and Emotion than between Pair and Cause. This measurable objective interference is consistent with the Pair F1 drop under joint training in Table 2, and motivates a pair-preserving schedule that refines utterance-level heads without perturbing the pair pathway.

F.3 Cross-Module Temporal Misalignment Diagnostic

We quantify whether **ATDG** aligns heterogeneous notions of recency across modules using a per-dialogue diagnostic.

Definition. ATDG maps label-free dialogue statistics $\psi(D)$ to a latent factor, which is projected to (i) modality-specific sequential decay rates $\boldsymbol{\lambda}(D) = \{\lambda_m(D)\}_{m=1}^{M}$ used in KS, and (ii) a graph decay rate $\gamma(D)$ used in SG. We summarize the sequential side by

$$\bar{\lambda}(D) \;=\; \frac{1}{M}\sum_{m=1}^{M} \lambda_m(D). \tag{21}$$

Because $\bar{\lambda}(D)$ and $\gamma(D)$ are strictly positive and live in different parameterizations, we compare them in log-space:

$$\Delta(D) \;=\; \big|\log \bar{\lambda}(D) - \log \gamma(D)\big| \;=\; \left|\log \frac{\bar{\lambda}(D)}{\gamma(D)}\right|. \tag{22}$$

A smaller $\Delta(D)$ indicates stronger cross-module temporal coherence, i.e., KS and SG imply more similar effective horizons for the same dialogue.

How we obtain $\bar{\lambda}(D)$ and $\gamma(D)$ for each method. We compute $\Delta(D)$ on the evaluation split under the following protocol.

- **ATDG+DP** and **ATDG-IndepPriors** (dialogue-conditioned): we use the per-dialogue decay outputs produced by ATDG, i.e., $\bar{\lambda}(D)$ and $\gamma(D)$.
- **ATDG-GlobalPriors** (dialogue-agnostic): we use the learned global (dialogue-independent) decay constants for KS and SG, and reuse the same $\bar{\lambda}$ and γ for all dialogues.
- **HiLo** (Li et al., 2025): since HiLo does not expose explicit decay parameters, we construct proxy horizons $H_{\text{seq}}(D) = T$ and $H_{\text{int}}(D) = \text{mean}\,|i-j|$ over (reply-adjacency∪ same-speaker) edges, and map them to proxy rates via $\bar{\lambda}(D) = 1/H_{\text{seq}}(D)$ and $\gamma(D) = 1/H_{\text{int}}(D)$.

We then report the distribution of $\Delta(D)$ across dialogues (Figure 4).

Non-degeneracy of learned decays. To verify that ATDG does not collapse to a constant prior, we visualize the marginal distributions of $\bar{\lambda}(D)$ and $\gamma(D)$ across dialogues for the final **ATDG+DP** model (Figure 9).

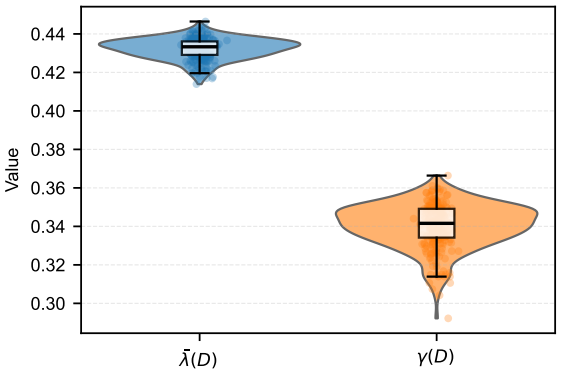

Figure 9: Marginal distributions of dialogue-conditioned decay parameters for the final **ATDG+DP**: mean sequential decay $\bar{\lambda}(D)$ and graph decay $\gamma(D)$. Non-collapsed distributions indicate that the learned time scale varies across dialogues.

Table 7: Pair extraction performance stratified by temporal distance, showing that ATDG+DP yields consistent gains at medium and long ranges where non-local reasoning is required (mean±std over 3 seeds).

Variant	Dist.*	F1	AUPRC	P	R
ATDG+DP	0	67.53±0.19	68.30±0.58	57.19±0.86	82.48±1.86
	[1, 2]	48.78±0.81	47.41±1.48	44.61±0.30	53.84±1.86
	[3, 6]	6.09±1.97	12.45±1.05	25.10±8.83	3.48±1.14
ATDG −IndepPriors	0	67.28±0.28	68.75±0.40	56.81±1.01	82.51±1.58
	[1, 2]	47.69±0.68	46.83±0.75	42.40±0.25	54.51±1.50
	[3, 6]	13.85±3.70	11.24±0.46	23.28±1.10	10.20±3.68
ATDG −GlobalPriors	0	67.23±0.63	68.53±0.43	58.41±1.41	79.24±1.46
	[1, 2]	47.63±0.35	48.48±0.56	42.72±1.30	53.94±2.77
	[3, 6]	13.18±3.79	12.24±1.28	24.04±4.44	9.70±3.95
ATDG+DP −w/o KS	0	67.89±0.62	68.82±0.64	58.99±0.82	80.01±2.44
	[1, 2]	46.08±0.96	46.56±3.13	45.14±4.83	47.69±4.33
	[3, 6]	7.04±2.41	11.23±0.26	21.36±4.63	4.23±1.55
ATDG+DP −w/o SG	0	67.90±0.47	68.42±0.96	60.58±1.10	77.32±2.81
	[1, 2]	47.60±1.56	46.35±0.31	44.90±1.69	50.67±2.06
	[3, 6]	8.31±2.54	12.26±0.92	23.75±4.27	5.22±1.97

* Dist. bins are defined by $\Delta = i - j$. The numbers of gold pairs for distances 0, [1, 2], and [3, 6] are 957, 694, and 134, respectively.

F.4 Distance Bin Performance

How hard is the long-tail in ECF? Table 7 reports ATDG+DP performance by temporal distance bin along with bin support. Performance is strong at Δ=0 and degrades substantially as distance increases, with the tail bin having very limited support. This quantifies a key challenge of ECF: long-range emotion triggers are rare and therefore difficult to learn reliably. Our main text therefore emphasizes not only average improvements but also stability under sparse regimes, where consistent priors across paths can reduce variance (Figure 5).

F.5 Sanity Perturbations of Dialogue Statistics

We denote by Clean (SAN-Base) the inference-time evaluation without perturbing $\psi(D)$, using the same checkpoint and decoding hyperparameters as in the main experiments. All sanity perturbations

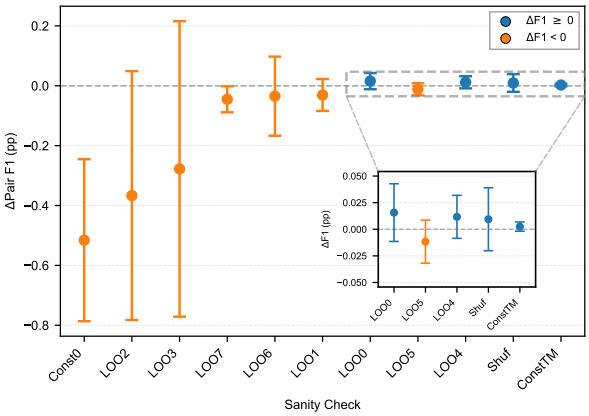

Figure 10: Eval-only sanity perturbations of dialogue statistics $\psi(D)$. ΔPair F1 (pp) relative to the clean setting shows stable behavior under mild perturbations and degradation under adversarial swaps.

Table 8: Sanity checks with the largest absolute ΔPair F1 (pp) on the test set (mean±std over 3 seeds).

Sanity	ΔPair F1	ΔAUPRC	ΔEmo F1	ΔCause F1
SAN-LOO2	-0.37±0.42	+0.01±0.01	+0.03±0.06	+0.03±0.06
SAN-LOO3	-0.28±0.49	+0.01±0.00	+0.02±0.03	-0.00±0.05
SAN-LOO6	-0.03±0.13	+0.00±0.01	+0.04±0.07	+0.02±0.04

are applied only at inference time, and we report ΔPair F1 (pp) relative to Clean.

Are ATDG priors overly sensitive to the guiding statistics? A natural concern is that ATDG might overfit to particular components of $\psi(D)$. We conduct inference-only sanity perturbations by shuffling or partially corrupting components of $\psi(D)$ at test time. As shown in Figure 10 and Table 8, performance changes remain small (sub-point ΔPair F1), suggesting that ATDG does not rely on a single fragile statistic and that consistent injection across KS and SG yields robust behavior.

Heavy-noise corruption via in-batch swap. To stress-test ATDG under stronger noise, we additionally apply a dialogue-level in-batch swap corruption: for each dialogue D in a mini-batch, with probability p we replace its statistics vector $\psi(D)$ with $\psi(D')$ from another dialogue in the same mini-batch via a random permutation, while keeping the dialogue inputs (x, a, v, s) unchanged. We consider $p \in \{0.1, 0.3, 0.5\}$ and evaluate the final **ATDG+DP** checkpoints (3 seeds). To avoid confounding from re-tuning decoding, we tune the pair threshold τ once on the clean validation split and reuse it unchanged for all corrupted test evaluations. Figure 11 reports ΔPair F1 (pp) relative to Clean (SAN-Base), and shows that performance remains

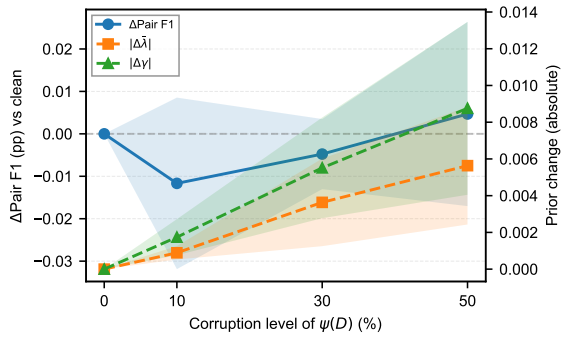

Figure 11: Robustness to inference-only corruption of $\psi(D)$ via in-batch swap. ΔPair F1 (pp) is reported relative to clean setting, showing graceful degradation as dialogue-pace statistics are corrupted.

Table 9: Data efficiency on the ECF test set under varying training fractions (mean±std over 3 seeds).

Train Fraction	Pair F1	Pair AUPRC	Emo. F1	Cause F1
25%	52.24±0.93	47.21±0.84	73.01±1.70	67.19±1.29
50%	55.13±0.62	52.33±0.82	77.98±0.56	70.70±0.51
75%	56.53±0.92	54.33±0.63	78.65±0.26	72.03±0.20
100%	57.92±0.52	55.58±0.57	79.33±0.30	73.26±0.80

stable up to 50% corruption, indicating that ATDG behaves as a robust inductive bias rather than a brittle shortcut.

Sanity check: the intervention changes ATDG decays. Even when Pair F1 is stable, the swap meaningfully perturbs ATDG outputs: the average per-dialogue changes in $\log \bar{\lambda}(D)$ (mean KS decay) and $\log \gamma(D)$ (SG decay) increase with p, confirming that the corruption substantially alters the mechanism-level time scales.

F.6 Missing Modality Robustness

Robustness to missing channels. Figure 12 evaluates robustness by randomly dropping modalities at test time and reporting the change in Pair F1 relative to the clean setting. As expected, performance decreases as the missing rate increases. Crucially, the degradation is gradual rather than catastrophic, suggesting that DP does not rely on a single modality-specific shortcut. Instead, temporal priors injected into KS/SG provide a modality-agnostic structural constraint that remains applicable even when some channels are absent.

F.7 Data Efficiency Numbers

Table 9 provides the exact data-efficiency numbers referenced in Figure 7. Pair extraction benefits most from additional data, which is consistent with structured link prediction requiring more supervision than utterance-level classification. Nevertheless, the strong 25% regime performance suggests that ATDG+DP provides a meaningful inductive bias when supervision is scarce.

Table 10: Sensitivity to the SG window size W on the ECF test set (mean±std over 3 seeds).

W	Pair F1	Pair AUPRC	Emo. F1	Cause F1
2	57.45±0.59	55.55±0.20	79.10±0.51	72.43±0.32
4	57.54±0.18	55.76±0.60	79.03±0.13	72.63±0.12
6	57.92±0.52	55.58±0.57	79.33±0.30	73.26±0.80
8	57.16±0.64	55.47±0.64	78.79±0.12	72.30±0.45

F.8 Window Size Sensitivity

Sensitivity to SG neighborhood size. Table 10 varies the SG window size W. Performance peaks at a moderate window (W=6) and degrades for both smaller and larger neighborhoods. This trend matches the intended role of SG as a local cross-speaker refiner: too small a window underutilizes short-range interaction cues, while too large a window propagates noisy and weakly-related messages that degrade SG-refined utterance representations and the auxiliary emotion/cause heads. Since the pair scorer consumes only KS representations, we keep SG local to improve utterance-level predictions without perturbing the pair pathway.

F.9 Emotion Per-Class Performance

Table 11 reports per-class emotion metrics for ATDG+DP. Performance is substantially higher on frequent classes (e.g., `neutral`, `joy`, `surprise`) than on rare classes (e.g., `fear`, `disgust`), which is consistent with class-imbalance effects in MELD/ECF. This result motivates reporting Macro-F1 in the main paper and suggests that future work may benefit from targeted imbalance mitigation or label-efficient adaptation for rare emotions.

F.10 Real-time Evaluation

Following the standard ECF benchmark protocol (Wang et al., 2023a), the main paper reports offline results to ensure fair comparison with prior work. We additionally report strict real-time evaluation to quantify performance under causal deployment constraints.

Protocol. In the real-time (online) setting, at step i the model may only access the observed prefix $D_{1:i} = \{u_1, \ldots, u_i\}$, and must not use any information from $u_{i+1:T}$. We obtain a strict online variant of ATDG+DP by enforcing causality in

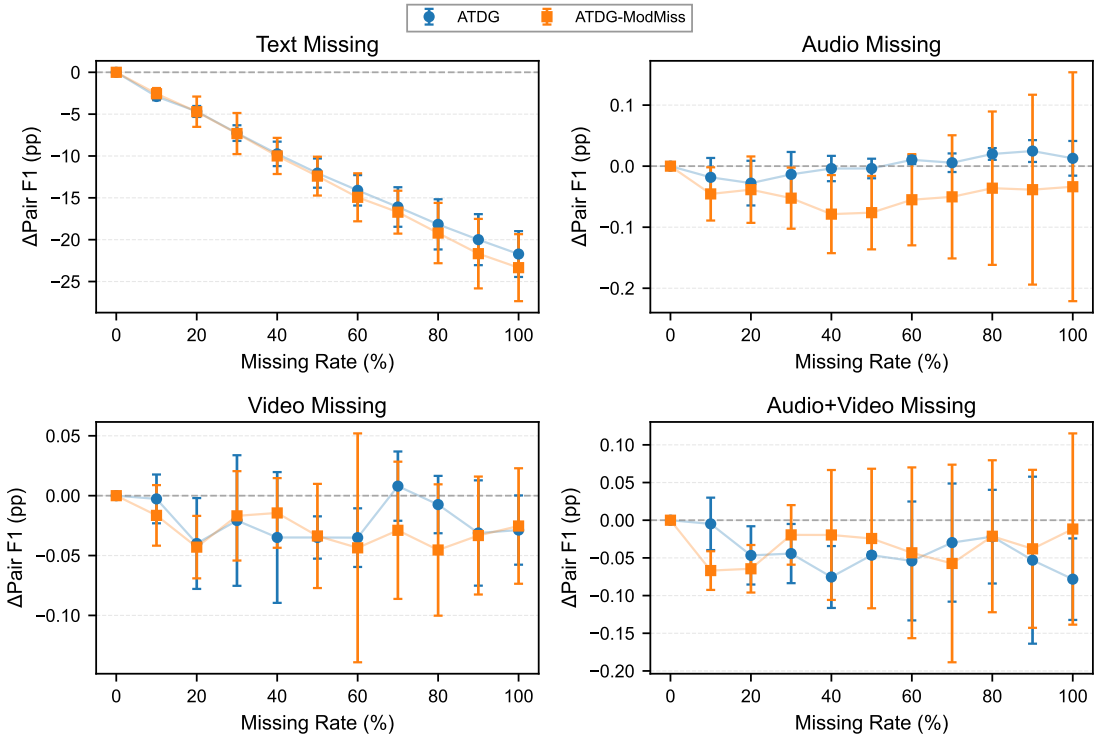

Figure 12: Robustness under missing modalities. ΔPair F1 (pp) is measured relative to the clean setting (0% missing), showing that dialogue-paced temporal priors maintain stable pair extraction as modality availability degrades.

Table 11: Per-class emotion performance for ATDG+DP on the test set (mean±std over 3 seeds), illustrating the effect of class imbalance across emotion categories.

Emotion	Precision	Recall	F1	Accuracy	AUPRC	Support
anger	50.24±2.07	48.85±2.55	49.51±1.93	87.07±0.53	50.22±0.63	333
disgust	10.00±17.32	1.27±2.19	2.25±3.89	96.80±0.07	9.82±0.75	79
fear	25.00±25.00	1.19±1.03	2.26±1.96	97.79±0.05	10.40±2.45	56
joy	55.02±0.79	62.94±2.92	58.68±0.93	85.19±0.22	63.00±1.57	429
neutral	73.43±0.85	73.15±1.72	73.27±0.46	76.70±0.08	78.77±0.65	1121
sadness	44.88±0.62	41.36±1.73	43.04±1.23	89.72±0.08	46.22±1.92	241
surprise	52.76±0.36	69.60±1.23	60.02±0.62	88.91±0.12	64.85±1.36	307

(i) dialogue-level representations and (ii) dialogue-conditioned temporal priors.

Causal representations. We encode each utterance independently with a pretrained encoder, $\mathbf{h}_i^{(0)} = \mathrm{Enc}(u_i)$, avoiding cross-utterance token leakage. We further replace bidirectional dialogue encoders with unidirectional ones (UniLSTM) for each modality stream, so the contextual state at time i depends only on the prefix: $\mathbf{h}_i = \mathrm{UniLSTM}(\mathbf{h}_{1:i}^{(0)})$. All multimodal fusion is applied on these causal states, and we report the same four modality settings as in Table 1.

Prefix-conditioned ATDG and online decoding. For ATDG, dialogue-pace statistics are computed from the prefix only, $\psi_i = \psi(D_{1:i})$, producing prefix-conditioned decays $\{\lambda_i^{(m)}\}_m$ for KS and γ_i for SG. We then apply the same DP injection as the static setting, except that both KS (Eq. 2) and SG (Eq. 4) use prefix-conditioned parameters at each step. For pair extraction, we keep the precedence constraint $j \le i$ and perform streaming decoding by scoring only available pairs (i, j) with the same biaffine scorer as Eq. (7), then aggregating step-wise decisions over $i = 1, \ldots, T$. All decoding hyperparameters (e.g., the threshold τ) are tuned on the validation split under the *same* online protocol and fixed for test, avoiding a static-online mismatch.

Table 12 summarizes online performance under the same four modality settings as Table 1, with all decoding hyperparameters tuned on the validation

Table 12: Real-time (online) evaluation of ATDG+DP on the ECF test set under strict prefix-only access (mean±std over 3 seeds).

Methods		Emotion Detection			Cause Detection			Pair Extraction		
		P	R	F1	P	R	F1	P	R	F1
ATDG+DP	Text-only	$79.54_{\pm 0.45}$	$79.64_{\pm 0.94}$	$79.59_{\pm 0.48}$	$66.49_{\pm 1.72}$	$76.01_{\pm 2.39}$	$70.89_{\pm 0.50}$	$50.53_{\pm 1.02}$	$62.42_{\pm 1.74}$	$55.83_{\pm 0.75}$
	+ Audio	$79.44_{\pm 0.75}$	$80.35_{\pm 1.28}$	$79.88_{\pm 0.32}$	$66.20_{\pm 2.70}$	$75.93_{\pm 3.77}$	$70.62_{\pm 0.16}$	$50.97_{\pm 2.24}$	$61.50_{\pm 4.25}$	$55.62_{\pm 0.69}$
	+ Video	$79.32_{\pm 0.75}$	$81.25_{\pm 1.28}$	$80.26_{\pm 0.44}$	$65.10_{\pm 0.86}$	$79.76_{\pm 1.95}$	$71.67_{\pm 0.36}$	$49.76_{\pm 1.54}$	$65.85_{\pm 2.71}$	$56.64_{\pm 0.38}$
	+ Audio + Video	$79.73_{\pm 0.50}$	$80.00_{\pm 0.88}$	$79.86_{\pm 0.39}$	$66.85_{\pm 1.44}$	$75.99_{\pm 1.74}$	$71.11_{\pm 0.65}$	$51.24_{\pm 1.14}$	$63.36_{\pm 2.28}$	$56.62_{\pm 0.33}$

split under the same online protocol. Compared with the offline protocol in Table 1, real-time evaluation yields only a modest degradation in pair extraction across modality settings. For Pair F1, the online scores drop by 1.05 (Text-only), 1.24 (+Audio), 0.43 (+Video), and 1.30 (+Audio+Video) absolute points, respectively. This gap is primarily driven by reduced access to future context in the strict prefix-only setting, which can suppress evidence accumulation for non-local triggers and delay confident link decisions. Overall, these results indicate that the proposed prefix-conditioned ATDG priors and KS/SG injection remain effective under strict deployment constraints, maintaining stable online performance without relying on future context.

F.11 Efficiency and Deployability in Real-Time

Beyond predictive quality, the practical viability of online MECPE hinges on predictable runtime under strict latency constraints, which are considered first-class requirements in latency-aware NLP and inference serving (Mendoza et al., 2024). Table 13 summarizes the model size, computational complexity, and runtime profile of our causal (prefix-only) ATDG+DP model. To ensure measurement stability, we report amortized latency (batch size 32) after a full-test warm-up pass to stabilize GPU kernels and caches, excluding CPU-side I/O overhead.

We observe an amortized per-sample latency of 28.90 ms with a 44.64 ms P95 tail latency and a throughput of 35.15 samples/s. Overall, the measured forward-pass latency lies in the lower end of the tens-to-hundreds-of-milliseconds response-latency budgets commonly targeted by user-facing serving systems (Mendoza et al., 2024).

Table 13: Model size and computational budget of our full real-time model (avg. over 3 seeds).

Metric	**ATDG+DP**
Params (M)	156.38
GFLOPs/sample	6.36
Latency Avg (ms)	28.90
Latency P50 (ms)	25.81
Latency P95 (ms)	44.64
Throughput (samples/s)	35.15

G AI Assistants

AI assistants were used solely to support language editing and clarity, including minor rephrasing and grammatical corrections. All technical content, modeling decisions, experimental design, results, and interpretations were conceived, implemented, and verified by the authors. No AI system was used to generate experimental results, analyze data, or make scientific claims.