

Sufficiency-Aware Adversarial Evidence Masking for Textual Emotion-Cause Pair Extraction in Conversations

Mengqiang Hu^{1,*} Qian Chen^{1,2,*†} Xin Guo¹ Suge Wang^{1,2}

¹School of Computer and Information Technology, Shanxi University, China

²Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, China

*Equal contribution. †Corresponding author.

leep87233@gmail.com chenqian@sxu.edu.cn

Abstract

Textual Emotion-Cause Pair Extraction in Conversations (TECPE) requires identifying emotion utterances, their categories, and the textual cause spans that explain them. Although recent systems improve aggregate scores, clean supervision can induce evidence bias: models exploit explicit emotion cues, cause–emotion proximity, and speaker/discourse priors rather than evidence sufficient for causal explanation. We introduce SAAM-EC, a sufficiency-aware adversarial evidence masking framework for span-grounded TECPE. SAAM-EC adversarially masks high-impact conversational evidence units during training, but treats masking as a semantic intervention rather than a label-preserving perturbation. When masking removes the cause span or necessary bridge context, the span may become unextractable or the relation unsupported. SAAM-EC therefore combines adversarial masking with a sufficiency gate and an extractability guard, training masked examples to recover the original span only when the remaining dialogue still supports the emotion–cause relation and preserves the annotated span; otherwise, it trains the model with an unanswerability objective. All auxiliary components are discarded at inference, leaving the backbone unchanged. Experiments on span-level TECPE and utterance-level ECPE benchmarks show consistent performance gains over clean backbones and masking-based alternatives. Diagnostics further show better recovery on distant and cross-speaker causes and stronger abstention under destructive masks. Our code is available at <https://github.com/0LeePriest/SAAM-EC>.

1 Introduction

Textual Emotion-Cause Pair Extraction in Conversations (TECPE) identifies all emotion–cause pairs in a dialogue, linking each emotion utterance and category to its textual cause span (Wang et al., 2024b; Li et al., 2023). Unlike utterance-level emotion recognition, TECPE must determine not only

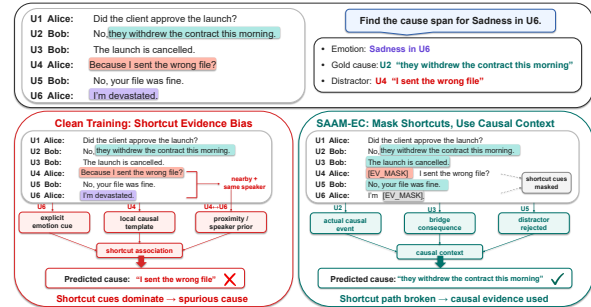


Figure 1: Evidence bias and shortcut masking in TECPE. Clean supervision may associate shortcut cues, such as explicit emotion words, local causal templates, proximity, or speaker priors, with a plausible but spurious cause span. SAAM-EC masks cue-bearing units during training to break these paths, encouraging reliance on causally grounded evidence: the triggering event, discourse bridge, and rejection of nearby distractors.

which emotion is expressed but also the textual evidence that accounts for it (Pereira et al., 2025). It also goes beyond utterance-level cause prediction by requiring span-level grounding, making the task inherently answerability-sensitive: most candidate emotion–cause utterance pairs contain no extractable answer, whereas valid pairs require precise causal evidence from multi-turn conversational context (Bhat and Modi, 2023; Zou et al., 2024).

Recent TECPE systems typically decompose the task into subtasks or cast cause extraction as reading comprehension over candidate utterances (Guo et al., 2024; Zhang et al., 2024; Ghahramani Kure et al., 2024). Although these models improve aggregate performance (Levchenko et al., 2024), clean supervision can still induce **evidence bias**: models exploit predictive conversational cues that are insufficient for span-grounded causal explanation (Du et al., 2024; Zhou et al., 2024). These cues include explicit emotion words, cause–emotion proximity, candidate-utterance priors, same-speaker histories, and local causal templates. While often useful, such cues become unre-

liable when causes are distant, cross-speaker, implicit, or mediated by intervening turns (Wang et al., 2024a; Ma et al., 2026). Models that rely on these shortcuts may therefore produce confident predictions even when the textual evidence required to justify the extracted span is absent (Yuan et al., 2024). Figure 1 illustrates this failure: a clean model may select the nearby same-speaker distractor, whereas SAAM-EC masks cue-bearing units to favor causal/discourse evidence.

A natural way to reduce shortcut reliance is to mask salient conversational evidence during training. However, directly applying standard adversarial or importance-based masking to span-grounded TECPE is unsafe (Lyu et al., 2024). Many supervised robustness methods implicitly assume label preservation, training perturbed examples toward the original target (Tran and Kretchmar, 2024). Evidence deletion in TECPE violates this assumption in many cases. Masking a peripheral utterance may preserve the original cause span, whereas masking the annotated span, candidate cause utterance, or a necessary bridge turn can make the answer unextractable from the remaining dialogue (Lavi et al., 2026). Forcing the model to recover the original start and end positions in such cases introduces false span supervision rather than robust learning. Therefore, a masked example should preserve the original target only when the remaining dialogue is semantically sufficient for the emotion–cause relation and the annotated span remains textually extractable. Otherwise, the target is NO_ANSWER: the annotated span is not extractable from the masked input, though the causal event may still be true.

We propose SAAM-EC, a sufficiency-aware adversarial evidence masking training framework for span-grounded TECPE. During training, SAAM-EC builds span-, utterance-, speaker-, and discourse-level evidence units for each positive query. An adversarial masker removes high-impact evidence units, while a sufficiency gate estimates whether the masked dialogue still supports the original emotion–cause relation. An extractability guard further blocks span supervision when the annotated cause span is no longer visible. Masked examples that remain sufficient and extractable train the backbone to recover the original span, whereas insufficient or non-extractable examples train the answerability head to predict NO_ANSWER. The masker, sufficiency gate, and teacher-based signals are used only during training, so SAAM-EC adds no inference-time modules beyond the un-

derlying TECPE backbone. Experiments on span-level TECPE and utterance-level ECPE benchmarks show consistent gains over clean backbones and masking-based alternatives; diagnostics show better recovery on distant/cross-speaker causes and stronger destructive-mask abstention. Our contributions are threefold:

- We identify evidence bias in span-grounded TECPE and formalize why label-preserving masking can introduce false supervision when evidence deletion changes span extractability.
- We formulate evidence masking as a sufficiency-aware semantic intervention problem and introduce SAAM-EC, which couples adversarial conversational evidence masking with sufficiency estimation and an extractability guard.
- We conduct controlled evaluations on span-level TECPE and utterance-level ECPE benchmarks, showing consistent gains over clean and masking-based baselines and more appropriate recovery and abstention behavior.

2 Related Work

Emotion-cause analysis identifies textual evidence explaining expressed emotions (Hu et al., 2024; Gu et al., 2024; Peng et al., 2025). Early emotion cause extraction conditions on a given emotion, whereas emotion-cause pair extraction jointly identifies emotions and causes (Xia and Ding, 2019; Yu et al., 2025). In conversations, this objective is complicated by speaker histories, multi-turn dynamics, cross-speaker dependencies, and implicit discourse relations (Xu et al., 2024; Shen et al., 2025; Zhang and Tan, 2025). Complementary work has introduced FOCUS, a Chinese customer-service dialogue summarization dataset with fine-grained aspect, sentiment, opinion, emotion, and customer-oriented summary annotations (Chen et al., 2026). Textual Emotion-Cause Pair Extraction in Conversations (TECPE) requires finer-grained supervision by requiring each prediction to include an emotion utterance, an emotion category, and a textual cause span, making it a span-grounded and answerability-sensitive task rather than utterance-level relation extraction. Recent TECPE models adopt pipeline, multi-stage, or machine reading comprehension formulations (Levchenko et al., 2024; Tian et al., 2025; Feng et al., 2026); they provide effective span-level backbones, but are typically optimized

on clean dialogues with standard supervision (Zhou et al., 2024; Yu et al., 2024; Hua et al., 2024).

Such training may reward shortcuts over sufficient causal evidence, including explicit emotion words, proximity between causes and emotions, priors over candidate utterances, same-speaker histories, and local causal templates (Xu et al., 2024; Shen et al., 2025; Zhang and Tan, 2025). These cues are predictive but not necessarily explanatory, especially for distant, cross-speaker, implicit, or discourse-mediated causes (Tian et al., 2025; Feng et al., 2026). Adversarial training and masking-based regularization improve robustness by perturbing inputs under label-preserving assumptions (Formento et al., 2024; Wang et al., 2025; Barkan et al., 2024), but span-grounded TECPE violates this assumption when masking removes the annotated span, the candidate cause utterance, or bridge context needed to justify the relation. SAAM-EC therefore treats masking as a semantic intervention whose target depends on sufficiency and extractability (Guo et al., 2025): if the masked dialogue still supports an extractable span, the model recovers it; otherwise, it learns NO_ANSWER, indicating that the original span is no longer extractable rather than that the causal event is false. This makes SAAM-EC complementary to existing TECPE backbones and distinct from label-preserving adversarial or masking-based training.

3 Methodology

Overview. We propose **SAAM-EC**, a training-time framework for textual emotion-cause pair extraction in conversations (TECPE). SAAM-EC keeps the test-time model identical to a question-answering (QA)-style TECPE backbone, but augments training with an adversarial evidence masker M_ϕ and a sufficiency gate S_ψ . The key idea is to treat evidence masking as a semantic intervention: after masking, the original span should be recovered only when the remaining dialogue is still sufficient and the annotated span remains extractable; otherwise, the masked query should be trained as NO_ANSWER. The masker, gate, and exponential moving average (EMA) teacher are discarded at inference. Figure 2 summarizes the framework.

3.1 Task and QA-style Backbone

A dialogue is $D = \{(s_i, u_i)\}_{i=1}^n$, where s_i is the speaker and $u_i = (w_{i,1}, \dots, w_{i,T_i})$ is an utterance. Let $\mathcal{E}$ be the emotion set and $\mathcal{E}_0 = \mathcal{E} \cup \{\text{NONE}\}$.

TECPE predicts

$$Y = \{(i, e_i, j, c_j)\}, \quad c_j = (w_{j,a}, \dots, w_{j,b}), \quad (1)$$

where u_i expresses emotion $e_i \in \mathcal{E}$ and span c_j in utterance u_j is its textual cause.

For each candidate emotion utterance u_i , emotion category e , and candidate cause utterance u_j , we form a query $q = q_{i,e,j}$ asking whether u_j contains a cause span for emotion e in u_i . The packed input $x_q = \text{Pack}(D, q)$ contains the full dialogue, speaker and utterance markers, an emotion-utterance marker around u_i , a candidate-cause marker around u_j , and the emotion query. The backbone encoder gives $H_q = E_\theta(x_q)$ and predicts

$$\begin{aligned} p_\theta^{\text{emo}}(e \mid D, i), \quad p_\theta^{\text{ans}}(z \mid D, q), \\ p_\theta^{\text{st}}(a \mid D, q), \quad p_\theta^{\text{end}}(b \mid D, q), \end{aligned} \quad (2)$$

where $z = 1$ means answerable and $z = 0$ denotes NO_ANSWER. Start and end distributions are restricted to tokens in u_j .

Let Q^+ be gold answerable queries and Q^- sampled unanswerable queries. For $q \in Q^+$, $z_q^* = 1$ with span (a_q^*, b_q^*) ; for $q \in Q^-$, $z_q^* = 0$. The clean loss is

$$\begin{aligned} \mathcal{L}_{\text{clean}} = - \sum_{i=1}^n \log p_\theta^{\text{emo}}(e_i^* \mid D, i) \\ + \sum_{q \in Q^+ \cup Q^-} \ell_{\text{qa}}(D, q), \end{aligned} \quad (3)$$

where

$$\begin{aligned} \ell_{\text{qa}}(D, q) = - \log p_\theta^{\text{ans}}(z_q^* \mid D, q) \\ + z_q^* [- \log p_\theta^{\text{st}}(a_q^* \mid D, q) \\ - \log p_\theta^{\text{end}}(b_q^* \mid D, q)]. \end{aligned} \quad (4)$$

Negative queries supervise answerability only. Non-emotion utterances are supervised by the NONE label when available and do not create positive cause-span queries.

3.2 Sufficiency-aware Evidence Intervention

For each positive query $q = q_{i,e,j}$, SAAM-EC builds conversational evidence units

$$\mathcal{G}_q = \{g_1, \dots, g_L\}. \quad (5)$$

These are semantic groups rather than isolated tokens, including the gold cause span, candidate cause utterance, emotion utterance, neighboring turns, same-speaker history, bridge turns, explicit emotion cues, and local causal templates. Cause

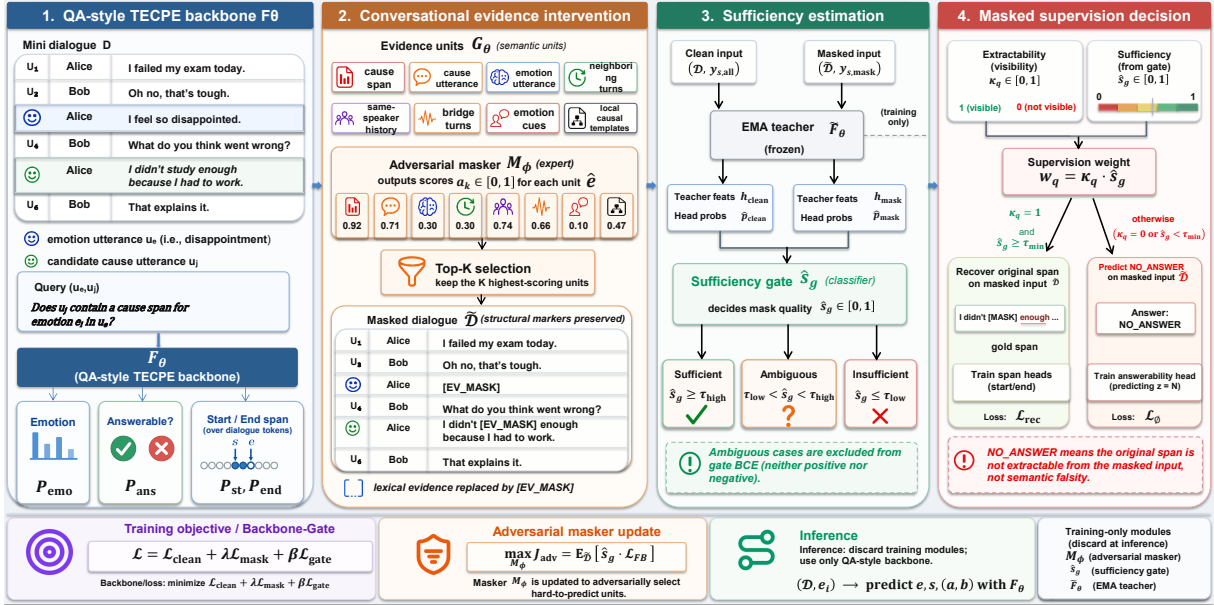


Figure 2: Overview of SAAM-EC: sufficiency-aware adversarial evidence masking for textual emotion-cause pair extraction in conversations.

spans and cause utterances are critical units because masking them can invalidate the original extractive target.

For each unit g_ℓ , the masker computes

$$\begin{aligned} r_\ell &= [\text{Pool}_{t \in g_\ell}(h_t); e_{\text{type}(g_\ell)}; f_\ell], \\ \alpha_\ell &= w_\phi^\top \tanh(W_\phi r_\ell + b_\phi), \end{aligned} \quad (6)$$

where h_t is the contextual token state, $e_{\text{type}(g_\ell)}$ is a learned unit-type embedding, and f_ℓ contains metadata such as turn distance, speaker match, unit length, and cue indicators. The masker selects exactly K units using a straight-through top- K (ST-TopK) operator:

$$\begin{aligned} \mathbf{z}_q &= \text{ST-TopK}_K(\alpha), \\ \mathbf{z}_q &\in \{0, 1\}^L, \quad \|\mathbf{z}_q\|_0 = K. \end{aligned} \quad (7)$$

Let $\mathcal{S}$ be unmaskable structural positions, including speaker, utterance, separator, query, emotion-utterance, and candidate-cause markers. The token mask is

$$m_{q,t} = \mathbb{I}[t \notin \mathcal{S}] \left[1 - \prod_{\ell=1}^L (1 - z_{q,\ell} \mathbb{I}[t \in g_\ell]) \right]. \quad (8)$$

The masked dialogue $\tilde{D} = \text{Mask}(D, \mathbf{m}_q)$ is produced by replacing masked lexical tokens with a single neutral evidence mask token:

$$\tilde{x}_{q,t} = \begin{cases} [\text{EVI_MASK}], & m_{q,t} = 1, \\ x_{q,t}, & m_{q,t} = 0. \end{cases} \quad (9)$$

We do not expose unit types to the backbone through mask tokens; unit-type information is used only by M_ϕ and S_ψ . Replacement, rather than deletion, preserves token alignment and span indices.

The sufficiency gate estimates whether the masked dialogue still supports the original pair:

$$\begin{aligned} s_q &= S_\psi(D, \tilde{D}, q, \mathbf{z}_q) \\ &= \sigma \left(w_\psi^\top \tanh(W_\psi g_q + b_\psi) \right), \end{aligned} \quad (10)$$

with

$$g_q = [h_0; \tilde{h}_0; |h_0 - \tilde{h}_0|; f_{\text{mask}}; e_T], \quad (11)$$

where h_0 and $\tilde{h}_0$ are clean and masked aggregate states, f_{mask} summarizes mask coverage, and e_T contains EMA-teacher answerability and span-overlap features.

After clean warm-up, we initialize an EMA teacher $F_{\bar{\theta}} \leftarrow F_\theta$ and update

$$\bar{\theta} \leftarrow \rho \bar{\theta} + (1 - \rho) \theta. \quad (12)$$

The teacher predicts answerability $\hat{z}_T = \arg \max_z p_{\bar{\theta}}^{\text{ans}}(z \mid \tilde{D}, q)$ and, when answerable, a span $\hat{c}_T(\tilde{D}, q)$. For $q \in Q^+$, define

$$\begin{aligned} \kappa_q &= \mathbb{I} \left[\sum_{t \in c_q} m_{q,t} = 0 \right], \\ \delta_q &= \mathbb{I} \left[\sum_{t \in c_q} (1 - m_{q,t}) = 0 \right], \end{aligned} \quad (13)$$

where c_q is the gold cause span of q . Thus $\kappa_q = 1$ means the full gold span remains visible, while

$\delta_q = 1$ means the full gold span or full candidate cause utterance has been removed. The weak sufficiency label is

$$s_q^* = \begin{cases} 1, & \kappa_q = 1 \wedge \hat{z}_T = 1 \\ & \wedge \text{IoU}(\hat{c}_T(\tilde{D}, q), c_q) > 0, \\ 0, & \delta_q = 1 \wedge \hat{z}_T = 0, \\ \perp, & \text{otherwise.} \end{cases} \quad (14)$$

Here, IoU denotes span intersection over union. This conservative rule labels only high-precision sufficient or insufficient cases; ambiguous masks are excluded from the gate binary cross-entropy (BCE) loss to avoid noisy binary targets.

3.3 Learning Objective

For a masked positive query, define

$$\tilde{s}_q = \begin{cases} s_q^*, & s_q^* \in \{0, 1\}, \\ \text{sg}(s_q), & s_q^* = \perp, \end{cases} \quad (15)$$

where $\text{sg}(\cdot)$ stops gradients. Since the backbone is extractive, semantic sufficiency is not enough: the original span can be supervised only when its annotated text remains visible. We therefore use the extractability-guarded weight

$$\omega_q = \kappa_q \tilde{s}_q. \quad (16)$$

Here NO_ANSWER means the original span is not extractable from the masked input, not that the causal event is false.

The masked recovery and no-answer losses are

$$\begin{aligned} \mathcal{L}_{\text{rec}}(\theta; \tilde{D}, q) &= -\log p_\theta^{\text{ans}}(1 | \tilde{D}, q) \\ &\quad - \log p_\theta^{\text{st}}(a_q^* | \tilde{D}, q) \\ &\quad - \log p_\theta^{\text{end}}(b_q^* | \tilde{D}, q), \end{aligned} \quad (17)$$

$$\mathcal{L}_\emptyset(\theta; \tilde{D}, q) = -\log p_\theta^{\text{ans}}(0 | \tilde{D}, q). \quad (18)$$

The masked objective is

$$\begin{aligned} \mathcal{L}_{\text{mask}} &= \mathbb{E}_{q \in Q^+} [\omega_q \mathcal{L}_{\text{rec}}(\theta; \tilde{D}, q) \\ &\quad + (1 - \omega_q) \mathcal{L}_\emptyset(\theta; \tilde{D}, q)]. \end{aligned} \quad (19)$$

The gate is trained only on non-ambiguous weak labels:

$$\mathcal{L}_{\text{gate}} = \frac{1}{\max(1, |\mathcal{Q}_g|)} \sum_{q \in \mathcal{Q}_g} \text{BCE}(s_q, s_q^*), \quad (20)$$

$$\mathcal{Q}_g = \{q \in Q^+ : s_q^* \neq \perp\}.$$

The backbone and gate minimize

$$\min_{\theta, \psi} \mathcal{L}_{\text{clean}} + \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \beta_{\text{gate}} \mathcal{L}_{\text{gate}}. \quad (21)$$

Here λ_{mask} controls the strength of masked supervision, while β_{gate} controls the strength of the auxiliary sufficiency-gate supervision. The masker is updated adversarially with θ and ψ fixed:

$$\begin{aligned} \max_{\phi} \quad \mathcal{J}_{\text{adv}}(\phi) &= \mathbb{E}_{q \in Q^+} [\text{sg}(\omega_q) \mathcal{L}_{\text{rec}}(\theta; \tilde{D}, q)], \\ \|\mathbf{z}_q\|_0 &= K. \end{aligned} \quad (22)$$

Thus the masker is rewarded for difficult masks only when the remaining input is both sufficient and extractable. Destructive masks train abstention through $\mathcal{L}_\emptyset$ but do not become successful adversarial examples. Since Eq. (7) fixes the unit budget and Eq. (22) down-weights destructive masks, we use no budget, entropy, or degeneracy regularizer.

3.4 Training and Inference

Training first warms up F_θ on $\mathcal{L}_{\text{clean}}$ and initializes $F_{\tilde{\theta}} \leftarrow F_\theta$. For each minibatch, we construct Q^+ and Q^- , build evidence units for $q \in Q^+$, select K units, construct $\tilde{D}$, assign s_q^* , compute ω_q , update (θ, ψ) and ϕ by Eqs. (21)–(22), and update the EMA teacher by Eq. (12). Appendix A.6 gives the algorithm.

At inference, M_ϕ , S_ψ , and $F_{\tilde{\theta}}$ are removed. For each utterance u_i , we keep non-NONE emotions satisfying $p_\theta^{\text{emo}}(e | D, i) \geq \tau_e$. For each retained e and candidate cause utterance u_j , we keep $q_{i,e,j}$ if $p_\theta^{\text{ans}}(1 | D, q_{i,e,j}) \geq \tau_z$. The span is decoded as

$$\begin{aligned} (\hat{a}, \hat{b}) &= \arg \max_{\substack{1 \leq a \leq b \leq T_j \\ b-a+1 \leq L_{\text{max}}}} p_\theta^{\text{st}}(a | D, q) \\ &\quad p_\theta^{\text{end}}(b | D, q). \end{aligned} \quad (23)$$

Thresholds τ_e, τ_z and optional L_{max} are selected on the development set and fixed for test evaluation.

4 Experiments

We evaluate SAAM-EC on span-level TECPE and utterance-level ECPE. For TECPE, we use ECF 2.0 from SemEval-2024 Task 3 Subtask 1 (Wang et al., 2024b) with UIC Two-Step QA (Chandakacherla et al., 2024) and THOR-ECAC (Rusnachenko and Liang, 2024). For ECPE, we use RECCON-DD/IE (Poria et al., 2021) under the ECPE-D-DD/ECPE-D-IE protocol with PRG-MoE (Jeong and Bak, 2023) and CENTER (Wang et al., 2024a); RECCON-DD is used for training and in-domain testing, and RECCON-IE only for cross-domain testing. We report official ECF weighted strict F1 (w-SF1) and weighted proportional F1

Method	w-SF1	w-PF1
UIC Two-Step QA	18.39	24.42
+SAAM-EC	19.62	26.27
Δ	+1.23	+1.85
THOR-ECAC	12.79	24.32
+SAAM-EC	13.74	25.56
Δ	+0.95	+1.24

Table 1: Span-level TECPE results on ECF 2.0. Δ denotes the absolute gain of SAAM-EC over each reproduced backbone; w-PF1 is the primary metric. Scores are percentages.

Method	RECCON-DD			RECCON-IE		
	P	R	F1	P	R	F1
PRG-MoE	58.95	55.67	57.26	51.95	20.02	28.90
+SAAM-EC	59.71	57.36	58.51	52.47	23.08	32.06
Δ	+0.76	+1.69	+1.25	+0.52	+3.06	+3.16
CENTER	64.94	60.24	62.50	43.55	67.50	52.94
+SAAM-EC	65.52	61.78	63.60	44.95	68.32	54.22
Δ	+0.58	+1.54	+1.10	+1.40	+0.82	+1.28

Table 2: Utterance-level ECPE results on RECCON. RECCON-DD and RECCON-IE evaluate in-domain and cross-domain generalization, respectively; pair F1 is the primary metric. Scores are percentages.

(w-PF1), using w-PF1 as primary, and RECCON pair-level precision, recall, and F1 (P/R/F1), using pair F1 as primary. Results are averaged over three seeds on a single NVIDIA A100; Appendix B gives backbone configurations, masked-target definitions, and hyperparameters.

4.1 Main Results

Tables 1 and 2 show controlled comparisons under fixed data splits, candidate construction, and inference decoders, so differences isolate the SAAM-EC training intervention. SAAM-EC improves every backbone-evaluation setting. On ECF 2.0, it raises w-PF1 by 1.85 points for UIC Two-Step QA and 1.24 for THOR-ECAC, with w-SF1 gains of 1.23 and 0.95; the strict-F1 gains indicate better span grounding rather than only partial-overlap recovery. On RECCON-DD, pair F1 improves by 1.25 with PRG-MoE and 1.10 with CENTER; on RECCON-IE, it improves by 3.16 and 1.28. The larger PRG-MoE gain on RECCON-IE is mainly recall-driven (+3.06 R vs. +0.52 P), suggesting better recovery of valid cross-domain pairs. Since the masker, sufficiency gate, and teacher are discarded at test time, these gains reflect training-time supervision, not extra inference.

4.2 Ablation Studies

Setup. Unless otherwise stated, ablations use UIC Two-Step QA on ECF 2.0 and PRG-MoE on RECCON-DD/IE. We report ECF w-SF1/w-PF1 and RECCON pair-level P/R/F1 as percentages. All masking variants share the same evidence-unit inventory, mask budget, replacement token, training schedule, and inference decoder, differing only in mask selection and masked-target assignment.

Masking ablation. Table 3 jointly evaluates mask selection and masked-target assignment. Force-label variants preserve the original span or positive-pair label after masking, whereas sufficiency-aware variants recover the original target only when the remaining input is sufficient and visible; otherwise they train NO_ANSWER or NO_PAIR. ImportantMask-Force ranks evidence units by the drop in the EMA-teacher gold-target score under leave-one-out masking and masks the top K units. Force-label masking yields only marginal gains, and AdvMask-Force remains below SAAM-EC despite using the same adversarial selector, showing that selection alone is insufficient. RandomMask-SAAM outperforms all force-label variants, identifying target reassignment as the dominant factor. SAAM-EC further improves over RandomMask-SAAM, indicating that adversarial selection is useful when coupled with sufficiency-aware supervision. On RECCON-IE, force-label variants mainly trade precision for recall, consistent with label noise from preserving positive targets after destructive masking.

Component ablation. Table 4 evaluates the core components of SAAM-EC. The extractability guard is most critical: removing it drops ECF w-PF1 and RECCON-DD F1 by 1.03 and 1.00 points, showing that original targets should be preserved only when the span or cause utterance remains visible. Removing the abstention-objective branch also hurts consistently, confirming that destructive masks should train abstention rather than be discarded. The learned gate outperforms rule-only assignment, whereas EMA-teacher features mainly stabilize gate decisions. Discourse evidence is also useful: removing bridge-turn and same-speaker units reduces RECCON-IE F1 by 1.29 points, especially under cross-domain shifts.

Sensitivity to masking strength. We vary the semantic mask budget K and masked-loss weight λ_{mask} in Figure 3. The default setting $K =$

Variant	Mask selector	Masked target	ECF		RECCON-DD			RECCON-IE		
			w-SF1	w-PF1	P	R	F1	P	R	F1
Clean Backbone	—	clean label	18.39	24.42	58.95	55.67	57.26	51.95	20.02	28.90
RandomMask-Force	random units	original target	18.51	24.70	58.72	56.18	57.42	50.87	20.78	29.51
ImportantMask-Force	teacher leave-one-out (LOO) units	original target	18.28	24.50	58.21	56.62	57.40	49.92	21.35	29.91
AdvMask-Force	adversarial units	original target	18.14	24.45	57.76	57.14	57.45	48.61	21.55	29.86
RandomMask-SAAM	random units	sufficiency-aware	19.08	25.79	59.18	56.89	58.01	51.83	22.35	31.23
SAAM-EC	adversarial units	sufficiency-aware	19.62	26.27	59.71	57.36	58.51	52.47	23.08	32.06

Table 3: Masking ablation over selector and target assignment. Force-label variants preserve original targets after masking, whereas SAAM variants use sufficiency-aware target reassignment.

Variant	Changed component	ECF		RECCON-DD			RECCON-IE		
		w-SF1	w-PF1	P	R	F1	P	R	F1
SAAM-EC	full model	19.62	26.27	59.71	57.36	58.51	52.47	23.08	32.06
w/o sufficiency gate	rule-only target assignment	19.24	25.87	59.34	56.82	58.05	51.74	22.32	31.19
w/o abstention objective	drop destructive masks; no abstention loss	18.86	25.36	58.68	56.74	57.69	49.61	22.68	31.13
w/o extractability guard	ignore span or cause-utterance visibility	18.57	25.24	58.23	56.81	57.51	48.73	22.81	31.07
w/o EMA features	gate uses contextual and mask features only	19.45	26.09	59.54	57.04	58.26	52.18	22.63	31.57
w/o bridge units	remove bridge-turn evidence units	19.20	25.82	59.12	56.72	57.90	51.88	22.18	31.07
w/o same-speaker units	remove same-speaker history units	19.36	25.96	59.39	56.83	58.08	52.30	22.46	31.42
w/o discourse units	remove bridge and same-speaker units	19.07	25.65	58.96	56.47	57.69	51.51	21.94	30.77

Table 4: Component ablation of SAAM-EC by removing gating, supervision, or evidence units.

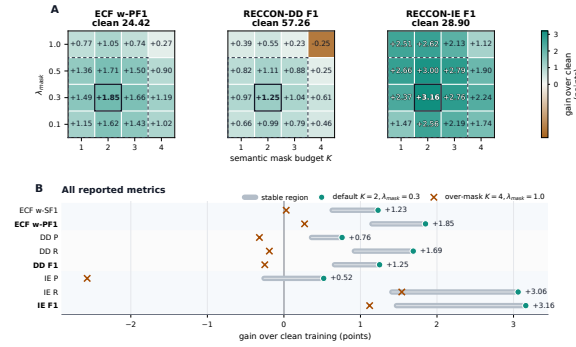


Figure 3: Hyperparameter sensitivity of SAAM-EC. Values show gains over clean training; ECF metrics are reported on the percentage scale. Panel A varies K and λ_{mask} , while Panel B compares stable, default, and over-masking regimes. Moderate masking is robust, whereas over-masking degrades performance.

$2, \lambda_{\text{mask}} = 0.3$ gives the best primary metrics, yielding +1.85 ECF w-PF1, +1.25 RECCON-DD F1, and +3.16 RECCON-IE F1 over clean training. Moderate settings are stable, whereas over-masking ($K = 4$ or $\lambda_{\text{mask}} = 1.0$) degrades performance.

4.3 Evidence Intervention Diagnostics

Ablations establish the value of sufficiency-aware target reassignment, but not whether models behave appropriately under explicit evidence deletion. We therefore conduct an ECF-only diagnostic with UIC Two-Step QA, inspecting span recovery when shortcuts are removed but the gold span remains,

and abstention when causal evidence is removed.

Intervention setup. For the clean backbone, AdvMask-Force, and SAAM-EC, we create two evaluation-only masks for each positive ECF query. A *sufficient mask* removes shortcut evidence, including explicit emotion cues, local causal templates, or nearby non-gold distractors, while keeping the gold span visible; its target remains the original span and recovery is measured by w-PF1. A *destructive mask* removes the full gold span or candidate cause utterance, making NO_ANSWER

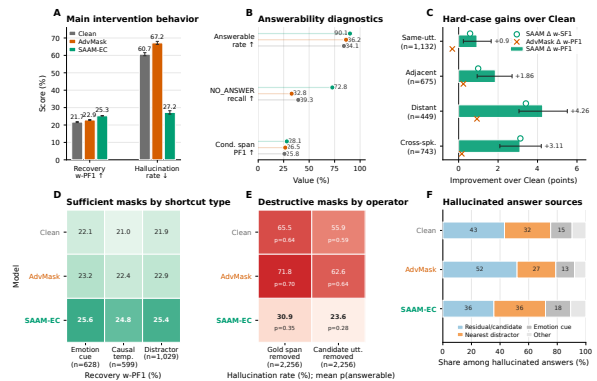


Figure 4: **Evidence-intervention diagnostics on ECF.** (A) Overall recovery under sufficient masks and hallucination under destructive masks. (B) Answerability diagnostics. (C) Hard-case gains. (D–F) Fine-grained recovery, destructive-mask hallucination, and hallucination-source breakdowns.

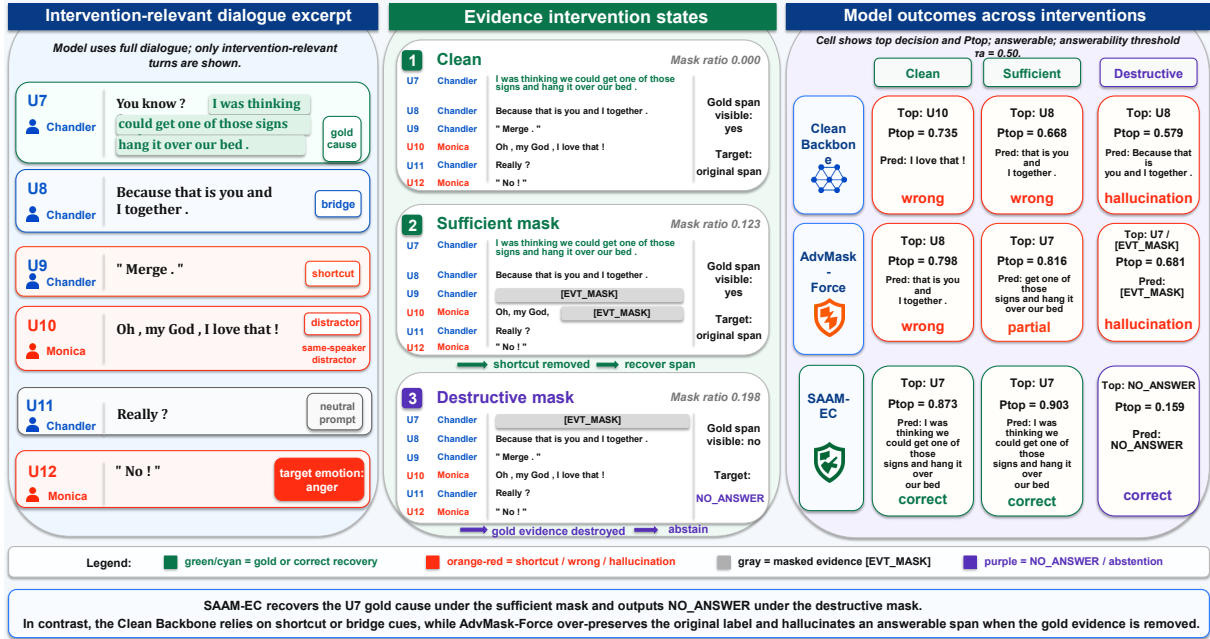


Figure 5: Qualitative evidence-intervention case. The sufficient mask removes shortcut/distractor cues while preserving the U7 gold span; the destructive mask removes U7 evidence, making NO_ANSWER the correct target. Each cell reports the top prediction and answerability score p_{top} ; the case visualization uses $\tau_a = \tau_z = 0.50$.

the correct target. Unsupported prediction is measured by the hallucination rate (HR), $HR = |\mathcal{D}_{des}|^{-1} \sum_{x \in \mathcal{D}_{des}} \mathbb{I}[\hat{z}(x) = 1]$, where $\mathcal{D}_{des}$ is the destructive-mask set and $\hat{z}(x) = 1$ denotes an answerable prediction; lower HR is better.

Overall intervention behavior. Figure 4A shows the desired asymmetry. Under sufficient masks, SAAM-EC reaches 25.3% recovery w-PF1, compared with 21.7% for the clean backbone and 22.9% for AdvMask-Force. Under destructive masks, it reduces hallucination from 60.7% and 67.2% to 27.2%. Figure 4B further shows improved answerability behavior: SAAM-EC answers 90.1% of recoverable inputs, achieves 28.1% conditional span PF1, and raises destructive-mask NO_ANSWER recall to 72.8%. Thus, the model does not simply answer more often; it better separates extractable from non-extractable inputs.

Sufficient versus destructive masks. Fine-grained results support the same conclusion. In Figure 4D, SAAM-EC improves recovery across all sufficient-mask types, including emotion cues, causal templates, and nearby distractors, indicating stronger use of the remaining causal context. Figure 4E shows the opposite effect for destructive interventions: AdvMask-Force hallucinates on 71.8% of span-removal cases and 62.6% of candidate-utterance-removal cases, both worse than the clean

backbone. SAAM-EC reduces these rates to 30.9% and 23.6%, confirming that destructive evidence deletion should train abstention rather than forced span recovery.

Hard cases and remaining errors. Figure 4C shows larger gains when shortcut cues are less reliable: +0.93/+1.86 w-PF1 for same-utterance/adjacent causes versus +4.26/+3.11 for distant/cross-speaker causes, where broader conversational evidence is required. Figure 4F further shows that SAAM-EC reduces residual candidate-location hallucinations relative to AdvMask-Force (35.7% vs. 51.9%); nearest non-gold distractors account for a similar 36.1% of SAAM-EC’s remaining hallucinations. Overall, the diagnostic confirms that evidence masking in TECPE is not inherently label-preserving: training should recover the original span only when the masked input remains sufficient and extractable, and abstain otherwise.

4.4 Case Study

Figure 5 shows an ECF 2.0 test case for the target emotion ANGER in U12. The gold cause is the distant cross-speaker span in U7: “I was thinking we could get one of those signs and hang it over our bed.” The case is challenging because U8 provides bridge context, U9 forms a local shortcut, and U10 is a nearby same-speaker affective distractor. All models receive the full dialogue; the figure shows

Setting	Backbone / stage	Train ratio	Latency	Inf. ratio	Extra params
ECF 2.0	UIC / SpanBERT	1.67×	146.2 / 146.5	1.00×	0
RECCON-DD/IE	PRG-MoE / pair cls.	1.62×	41.8 / 42.0	1.00×	0

Table 5: Efficiency of SAAM-EC relative to clean training. Train ratio takes clean training as 1.00×; latency is reported as clean/+SAAM-EC milliseconds per dialogue; extra parameters are inference-time parameters.

only intervention-relevant turns.

The clean backbone relies on local evidence: it selects U10 on the clean input ($p_{\text{top}} = 0.735$), shifts to U8 under the sufficient mask ($p_{\text{top}} = 0.668$), and still predicts an answerable U8 span after destructive masking ($p_{\text{top}} = 0.579$). AdvMask-Force moves toward the gold candidate under the sufficient mask, selecting U7 ($p_{\text{top}} = 0.816$), but recovers an incomplete span and hallucinates an answer at the masked U7 location under the destructive mask ($p_{\text{top}} = 0.681$). In contrast, SAAM-EC predicts the correct U7 span on the clean and sufficient inputs ($p_{\text{top}} = 0.873/0.903$) and outputs NO_ANSWER after destructive masking, lowering answerability to 0.159. This case illustrates the intended behavior of sufficiency-aware supervision: recover causes when sufficient evidence remains and abstain when extractive evidence is removed.

4.5 Efficiency

Table 5 shows that SAAM-EC increases training time but leaves deployment effectively unchanged. Training takes $1.67\times$ as long on ECF 2.0 with UIC and $1.62\times$ as long on RECCON with PRG-MoE, due to masked inputs, gate updates, adversarial masking, and EMA-teacher scoring. At inference, the masker, sufficiency gate, and EMA teacher are discarded, so SAAM-EC adds no parameters and keeps latency nearly identical: 146.2/146.5 ms on ECF and 41.8/42.0 ms on RECCON for clean/SAAM-EC-trained models.

5 Conclusion

We presented SAAM-EC, a sufficiency-aware adversarial evidence masking framework for span-grounded TECPE. By combining adversarial masking with a sufficiency gate and an extractability guard, SAAM-EC trains span recovery only when the original cause remains supported and visible, and otherwise trains abstention. Experiments on span-level TECPE and utterance-level ECPE benchmarks show consistent gains over

strong backbones without adding inference-time modules, especially for long-distance and cross-speaker causes. Overall, these results indicate that aligning masked supervision with both semantic sufficiency and span extractability leads to more robust evidence-grounded emotion-cause extraction.

Limitations

This work evaluates SAAM-EC on established text-based conversational emotion-cause benchmarks, covering span-level extraction on ECF 2.0 and utterance-level classification on RECCON-DD/IE with multiple representative backbones. These benchmarks provide a controlled and widely used setting for studying evidence-grounded TECPE/ECPE. A natural next step is to examine whether the same sufficiency-aware training objective brings similar benefits in broader dialogue settings, such as multilingual conversations or multimodal interactions that include acoustic and visual cues.

Our formulation follows the supervision available in standard emotion-cause datasets: textual cause spans or cause utterances are treated as the target evidence for an expressed emotion. This keeps the method aligned with benchmark definitions and avoids introducing additional annotation assumptions. Future work may explore richer discourse structures or domain-specific evidence units, while the present study intentionally uses a compact and interpretable evidence inventory that can be applied to different backbones without changing their inference architectures.

Ethical Considerations

This work studies text-based emotion-cause extraction in conversations using publicly released research benchmarks cited in the paper. We do not collect new private conversations, solicit new human annotations, or attempt to identify speakers beyond the information already present in the benchmark data. Our experiments follow the established train/dev/test protocols of the datasets, and the method is evaluated offline as a research contribution rather than as a deployed affective-computing system.

The task concerns emotions and their textual causes, which can be sensitive in real-world applications. Predictions produced by TECPE or ECPE systems should not be interpreted as definitive psychological judgments about speakers, nor should

they be used for high-stakes decisions such as hiring, education, clinical assessment, law enforcement, or workplace monitoring without informed consent, domain validation, and human oversight. SAAM-EC explicitly encourages abstention when the original evidence is no longer extractable, but this mechanism does not eliminate all unsupported or biased predictions.

The datasets and labels used for emotion-cause analysis may reflect the linguistic, cultural, and annotator-specific assumptions under which they were created. Models trained on these resources can therefore inherit biases in emotion expression, speaker behavior, dialogue genre, or cause attribution. We report controlled benchmark results and diagnostic analyses to characterize model behavior, but we do not claim that the learned associations are universal across languages, cultures, demographic groups, or interaction settings. Users adapting the method to new domains should examine dataset composition, annotation policy, error patterns, and group-level impacts before deployment.

SAAM-EC adds computation during training due to masked inputs, sufficiency estimation, and teacher scoring, but all auxiliary components are removed at inference. The experiments are run on a single NVIDIA A100 GPU, and the method does not require external LLM-based verification or user-facing text generation. We cite the original datasets and backbones, use them for their intended research purposes, and document implementation details to support reproducibility while respecting the licenses and usage terms of the underlying resources.

Acknowledgments

This work is supported by the Joint Key Project of the National Natural Science Foundation of China (No. U24A20335) and the National Natural Science Foundation of China (No. 62376143).

References

- Oren Barkan, Yonatan Toib, Yehonatan Elisha, Jonathan Weill, and Noam Koenigstein. 2024. [LLM explainability via attributive masking learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9522–9537, Miami, Florida, USA. Association for Computational Linguistics.
- Ashwani Bhat and Ashutosh Modi. 2023. [Multi-task learning framework for extracting emotion cause span and entailment in conversations](#). In *Proceedings of the 1st Transfer Learning for Natural Language Processing Workshop*, volume 203 of *Proceedings of Machine Learning Research*, pages 33–51. PMLR.
- Sharad Chandakacherla, Vaibhav Bhargava, and Natalie Parde. 2024. [UIC NLP GRADS at SemEval-2024 task 3: Two-step disjoint modeling for emotion-cause pair extraction](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1373–1379, Mexico City, Mexico. Association for Computational Linguistics.
- Qian Chen, Mengqiang Hu, and Xin Guo. 2026. [FOCUS: A fine-grained customer-oriented sentiment dialogue summarization dataset for Chinese customer service](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 22744–22764, San Diego, California, United States. Association for Computational Linguistics.
- Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2024. [Shortcut learning of large language models in natural language understanding](#). *Communications of the ACM*, 67(1):110–120.
- Jun-Hao Feng, Xia-Bing Zhou, Wen-Liang Chen, and Min Zhang. 2026. [Learning contextual information and task alignment for emotion cause extraction in conversation](#). *Frontiers of Computer Science*, 20(1):2001308.
- Brian Formento, Wenjie Feng, Chuan-Sheng Foo, Anh Tuan Luu, and See-Kiong Ng. 2024. [SemRoDe: Macro adversarial training to learn representations that are robust to word-level attacks](#). 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 8005–8028, Mexico City, Mexico. Association for Computational Linguistics.
- Alireza Ghahramani Kure, Mahshid Dehghani, Mohammad Mahdi Abootorabi, Nona Ghazizadeh, Seyed Arshan Dalili, and Ehsaneddin Asgari. 2024. [AIMA at SemEval-2024 task 3: Simple yet powerful emotion cause pair analysis](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1698–1703, Mexico City, Mexico. Association for Computational Linguistics.
- Xue Gu, Zhihan Zhou, Ziyao Meng, Jian Li, Tiago Gomes, Adriano Tavares, and Hao Xu. 2024. [EmoPrompt-ECPE: Emotion knowledge-aware prompt-tuning for emotion-cause pair extraction](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5678–5688, Torino, Italia. ELRA and ICCL.
- Hongyu Guo, Xueyao Zhang, Yiyang Chen, Lin Deng, and Binyang Li. 2024. [UIR-ISC at SemEval-2024 task 3: Textual emotion-cause pair extraction in conversations](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 770–776, Mexico City, Mexico. Association for Computational Linguistics.

- Siyi Guo, Myrl G. Marmarelis, Fred Morstatter, and Kristina Lerman. 2025. [Estimating causal effects of text interventions leveraging LLMs](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 170–190, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.
- Guimin Hu, Zhihong Zhu, Daniel Hershcovich, Lijie Hu, Hasti Seifi, and Jiayuan Xie. 2024. [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.
- Yuncheng Hua, Yujin Huang, Shuo Huang, Tao Feng, Lizhen Qu, Chris Bain, Richard Bessed, and Gholamreza Haffari. 2024. [Causal discovery inspired unsupervised domain adaptation for emotion-cause pair extraction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8139–8156, Miami, Florida, USA. Association for Computational Linguistics.
- DongJin Jeong and JinYeong Bak. 2023. [Conversational emotion-cause pair extraction with guided mixture of experts](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3288–3298, Dubrovnik, Croatia. Association for Computational Linguistics.
- Maor Juliet Lavi, Tova Milo, and Mor Geva. 2026. [Detecting \(un\)answerability in large language models with linear directions](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–699, Rabat, Morocco. Association for Computational Linguistics.
- Sofia Levchenko, Rafał Wolert, and Piotr Andrzejewicz. 2024. [PWEITINLP at SemEval-2024 task 3: Two step emotion cause analysis](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1097–1105, Mexico City, Mexico. Association for Computational Linguistics.
- Wei Li, Yang Li, Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria. 2023. [ECPEC: Emotion-cause pair extraction in conversations](#). *IEEE Transactions on Affective Computing*, 14(3):1754–1765.
- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. [Towards faithful model explanation in NLP: A survey](#). *Computational Linguistics*, 50(2):657–723.
- Tianxiang Ma, Weijie Feng, Xinyu Wang, and Zhiyong Cheng. 2026. [Emotion-cause pair extraction in conversations via semantic decoupling and alignment](#). In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 12697–12711, San Diego, California, United States. Association for Computational Linguistics.
- Sancheng Peng, Lihong Cao, Guojun Wang, Zhouhao Ouyang, Yongmei Zhou, and Shui Yu. 2025. [A survey on textual emotion cause extraction in social networks](#). *Digital Communications and Networks*, 11(2):524–536.
- Patrícia Pereira, Helena Moniz, and João Paulo Carvalho. 2025. [Deep emotion recognition in textual conversations: A survey](#). *Artificial Intelligence Review*, 58(1):10.
- Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. [Recognizing emotion cause in conversations](#). *Cognitive Computation*, 13(5):1317–1332.
- Nicolay Rusnachenko and Huizhi Liang. 2024. [nicolay-r at SemEval-2024 task 3: Using flan-t5 for reasoning emotion cause in conversations with chain-of-thought on emotion states](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 22–27, Mexico City, Mexico. Association for Computational Linguistics.
- Zhiyu Shen, Yunhe Pang, Yanghui Rao, and Jianxing Yu. 2025. [CoE: A clue of emotion framework for emotion recognition in conversations](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23548–23563, Vienna, Austria. Association for Computational Linguistics.
- Yuanhe Tian, Pengsen Cheng, Fei Xia, Jiayong Liu, Yongdong Zhang, and Yan Song. 2025. [Emotion cause extraction in conversations with response graphing](#). In *Natural Language Processing and Chinese Computing*, volume 15363 of *Lecture Notes in Computer Science*, pages 69–81. Springer, Singapore.
- Son Quoc Tran and Matt Kretchmar. 2024. [Towards robust extractive question answering models: Rethinking the training methodology](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2222–2236, Miami, Florida, USA. Association for Computational Linguistics.
- 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. Association for Computational Linguistics.
- 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.
- Yang Wang, Chenghao Xiao, Yizhi Li, Stuart E. Middleton, Noura Al Moubayed, and Chenghua Lin. 2025. Adversarial defense without adversarial defense: Enhancing language model robustness via instance-level principal component removal. *Transactions of the Association for Computational Linguistics*, 13:1381–1409.
- 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.
- Bo Xu, Longjiao Li, Wei Luo, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2024. Beyond linguistic cues: Fine-grained conversational emotion recognition via belief-desire modelling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2318–2328, Torino, Italia. ELRA and ICCL.
- Yang Yu, Xin Alex Lin, Changqun Li, Shizhou Huang, and Liang He. 2024. MGCL: Multi-granularity clue learning for emotion-cause pair extraction via cross-grained knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1897–1907, Miami, Florida, USA. Association for Computational Linguistics.
- Zhaoxin Yu, Xinglin Xiao, and Wenji Mao. 2025. 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.
- Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do LLMs overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200, Miami, Florida, USA. Association for Computational Linguistics.
- Shen Zhang, Haojie Zhang, Jing Zhang, Xudong Zhang, Yimeng Zhuang, and Jinting Wu. 2024. Samsung research china-beijing at SemEval-2024 task 3: A multi-stage framework for emotion-cause pair extraction in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 536–546, Mexico City, Mexico. Association for Computational Linguistics.
- Tao Zhang and Zhenhua Tan. 2025. 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)*, pages 2064–2077, Vienna, Austria. Association for Computational Linguistics.
- Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614, Miami, Florida, USA. Association for Computational Linguistics.
- Jiajun Zou, Yexuan Zhang, Sixing Wu, Jinshuai Yang, Xuanmei Qin, Lizhi Ying, Minghu Jiang, and Yongfeng Huang. 2024. A machine reading comprehension framework for recognizing emotion cause in conversations. *Knowledge-Based Systems*, 289:111532.

A Additional Method Details

A.1 Query Construction

For a dialogue with gold tuples Y , positive queries are

$$Q^+(D) = \{q_{i,e,j} : (i, e, j, c_j) \in Y\}. \quad (24)$$

Each positive query stores its gold span $c_q = c_j$ and boundaries (a_q^*, b_q^*) . For each gold emotion instance (i, e) , define its cause-utterance set

$$C_{i,e} = \{j : (i, e, j, c_j) \in Y\}. \quad (25)$$

Negative queries are sampled from

$$\mathcal{N}_{i,e} = \{q_{i,e,j} : j \in \{1, \dots, n\}, j \notin C_{i,e}\}. \quad (26)$$

We sample at most $r|C_{i,e}|$ negatives per emotion instance. The sampler prioritizes hard negatives from nearby turns and same-speaker turns, then samples uniformly from the remaining candidates. If the candidate pool is small, we use all negatives. Negative queries supervise answerability only.

At inference, the model enumerates all candidate cause utterances for each predicted non-NONE emotion unless a dataset-specific pruning rule is explicitly stated. If multiple predicted spans share the same (i, e, j) , we keep the highest-scoring span. The optional tuple score is

$$S(i, e, j, \hat{c}) = p_\theta^{\text{emo}}(e \mid D, i) p_\theta^{\text{ans}}(1 \mid D, q) p_\theta^{\text{st}}(\hat{a} \mid D, q) p_\theta^{\text{end}}(\hat{b} \mid D, q). \quad (27)$$

A.2 Conversational Evidence Units

Evidence units are constructed only for positive queries used in masked training. Duplicate units are merged, structural markers are excluded from masking, and overlapping lexical tokens are masked by union as in Eq. (8). The default inventory is shown in Table 6.

The budget K is a semantic-unit budget rather than a token budget. This is intentional: SAAM-EC intervenes on conversational evidence groups. To reduce hidden length bias, unit length and coverage statistics are included in f_ℓ and f_{mask} , and all masking baselines use the same evidence-unit inventory and K .

A.3 Straight-through Top-K and Embedding-level Masking

The forward pass of Eq. (7) uses hard top- K :

$$z_{q,\ell}^{\text{hard}} = \mathbb{I}[\alpha_\ell \text{ is among the top } K \text{ scores}]. \quad (28)$$

Unit type	Construction
Cause span	Gold span c_q for $q \in Q^+$; the most specific critical unit.
Cause utterance	All lexical tokens in candidate cause utterance u_j .
Emotion utterance	All lexical tokens in u_i , excluding structural markers.
Emotion cues	High-precision emotion lexicon matches or cue phrases, especially in u_i .
Neighbors	Existing turns in $\{u_{i-1}, u_{i+1}, u_{j-1}, u_{j+1}\}$.
Same-speaker history	Previous utterances u_k with $k < i$ and $s_k = s_i$, clipped to the nearest B turns.
Bridge turns	Each utterance u_k with $\min(i, j) < k < \max(i, j)$; optionally an aggregate bridge unit.
Causal templates	Local spans around cue patterns such as <i>because</i> , <i>since</i> , <i>when</i> , and event predicates.

Table 6: Default conversational evidence units for a positive query $q_{i,e,j}$.

For back-propagation, we use a differentiable relaxation

$$\bar{z}_q = \text{SoftTopK}_K(\alpha; \tau_{\text{ST}}), \quad (29)$$

and the straight-through estimator

$$\mathbf{z}_q^{\text{ST}} = \mathbf{z}_q^{\text{hard}} + \bar{\mathbf{z}}_q - \text{sg}(\bar{\mathbf{z}}_q). \quad (30)$$

The hard and relaxed token masks are

$$m_{q,t}^{\text{hard}} = \mathbb{I}[t \notin \mathcal{S}] \left[1 - \prod_{\ell=1}^L (1 - z_{q,\ell}^{\text{hard}} \mathbb{I}[t \in g_\ell]) \right], \quad (31)$$

$$\bar{m}_{q,t} = \mathbb{I}[t \notin \mathcal{S}] \left[1 - \prod_{\ell=1}^L (1 - \bar{z}_{q,\ell} \mathbb{I}[t \in g_\ell]) \right]. \quad (32)$$

During the masker update, hard replacement is implemented with an embedding-level straight-through mask:

$$m_{q,t}^{\text{ST}} = m_{q,t}^{\text{hard}} + \bar{m}_{q,t} - \text{sg}(\bar{m}_{q,t}), \quad (33)$$

$$\tilde{\mathbf{e}}_{q,t} = (1 - m_{q,t}^{\text{ST}}) \mathbf{e}_{q,t} + m_{q,t}^{\text{ST}} \mathbf{e}_{\text{mask}}.$$

Thus the forward input corresponds to hard neutral-token replacement, while gradients to M_ϕ follow the soft relaxation. The temperature τ_{ST} is annealed from 1.0 to 0.5 after warm-up; all compared variants use the same schedule.

A.4 Gate Features and Weak Labels

The mask metadata vector is

$$f_{\text{mask}} = [r_c; r_u; r_D; b_{\text{type}}; |i - j|; \mathbb{I}[s_i = s_j]], \quad (34)$$

where

$$\begin{aligned} r_c &= \frac{1}{|c_q|} \sum_{t \in c_q} (1 - m_{q,t}), \\ r_u &= \frac{1}{|u_j|} \sum_{t \in u_j} (1 - m_{q,t}), \\ r_D &= \frac{1}{|D|} \sum_{t \in D} m_{q,t}, \end{aligned} \quad (35)$$

and b_{type} is a histogram over selected unit types. The EMA-teacher feature vector is

$$e_T = [p_T^c; p_T^m; p_T^c - p_T^m; o_T; \ell_T^c - \ell_T^m], \quad (36)$$

where

$$p_T^c = p_{\theta}^{\text{ans}}(1 \mid D, q), \quad (37)$$

$$p_T^m = p_{\theta}^{\text{ans}}(1 \mid \tilde{D}, q), \quad (38)$$

$$o_T = \mathbb{I}[\hat{z}_T = 1] \text{IoU}(\hat{c}_T(\tilde{D}, q), c_q), \quad (39)$$

$$\ell_T^c = \log p_{\bar{\theta}}(y_q^* \mid D, q), \quad (40)$$

$$\ell_T^m = \log p_{\bar{\theta}}(y_q^* \mid \tilde{D}, q). \quad (41)$$

For $y_q^* = (z = 1, a_q^*, b_q^*)$,

$$\begin{aligned} \log p_{\bar{\theta}}(y_q^* \mid D, q) &= \log p_{\bar{\theta}}^{\text{ans}}(1 \mid D, q) \\ &\quad + \log p_{\bar{\theta}}^{\text{st}}(a_q^* \mid D, q) \\ &\quad + \log p_{\bar{\theta}}^{\text{end}}(b_q^* \mid D, q), \end{aligned} \quad (42)$$

with the analogous definition on $\tilde{D}$. These features are inputs to the gate only. If a minibatch contains no non-ambiguous weak labels, $\mathcal{L}_{\text{gate}}$ is set to zero for that minibatch.

A.5 Why the Extractability Guard is Necessary

A masked dialogue may still imply that an event caused an emotion, but an extractive QA head cannot recover the original textual span if some gold span tokens have been replaced. Applying start/end supervision in this case creates label noise. The guard

$$\omega_q = \kappa_q \tilde{s}_q \quad (43)$$

therefore enforces a stricter condition: span recovery is trained only when the original annotated span remains fully visible and the masked dialogue is judged sufficient. When $\kappa_q = 0$, the masked query is trained as NO_ANSWER for the extractive head, meaning that the original span is no longer extractable from the masked input.

A.6 Training Algorithm

Algorithm 1 SAAM-EC Training

```

1: Input: training dialogues, backbone  $F_{\theta}$ , masker  $M_{\phi}$ , gate
    $S_{\psi}$ , budget  $K$ , weights  $\lambda_{\text{mask}}, \beta_{\text{gate}}$ , EMA decay  $\rho$ .
2: Warm up  $F_{\theta}$  on  $\mathcal{L}_{\text{clean}}$  for  $N_{\text{warm}}$  steps.
3: Initialize  $F_{\bar{\theta}} \leftarrow F_{\theta}$ .
4: for each minibatch do
5:   Build  $Q^+$  and  $Q^-$ ; compute  $\mathcal{L}_{\text{clean}}$  on clean inputs.
6:   for each  $q = (i, e, j) \in Q^+$  do
7:     Construct  $\mathcal{G}_q$  and score units with  $M_{\phi}$ .
8:     Select  $K$  units with ST-TopK and construct  $\tilde{D}$  by
       neutral replacement masking.
9:     Run  $F_{\bar{\theta}}$  on  $(D, q)$  and  $(\tilde{D}, q)$ .
10:    Assign  $s_q^* \in \{0, 1, \perp\}$  by Eq. (14).
11:    Compute  $s_q, \tilde{s}_q$  by Eq. (15), and  $\omega_q$  by Eq. (16).
12:   end for
13:   Update  $(\theta, \psi)$  by minimizing Eq. (21), with  $\phi$  fixed.
14:   Update  $\phi$  by maximizing Eq. (22), with  $(\theta, \psi)$  fixed.
15:   Update EMA teacher:  $\bar{\theta} \leftarrow \rho \bar{\theta} + (1 - \rho) \theta$ .
16: end for
17: Inference: discard  $M_{\phi}, S_{\psi}$ , and  $F_{\bar{\theta}}$ ; use only  $F_{\theta}$ .

```

During the masker update, backbone and gate parameters are frozen. During the backbone/gate update, the sampled hard mask is treated as fixed. This avoids second-order optimization and keeps training comparable to standard adversarial data augmentation.

A.7 Hyperparameters and Controlled Variants

The main method-level hyperparameters are the semantic masking budget K , the masked-loss weight λ_{mask} , and the gate-loss weight β_{gate} . The EMA decay ρ , negative-sampling ratio r , warm-up length N_{warm} , SoftTopK temperature τ_{ST} , maximum span length L_{max} , and decoding thresholds τ_e, τ_z are selected on the development set and fixed for test evaluation. Unless stated otherwise, the main model uses no entropy regularizer, budget penalty, degeneracy penalty, LLM verifier, or adversarial parameter perturbation.

Variant	Purpose
Backbone	Clean QA-style TECPE baseline.
Random unit masking	Tests whether generic masking suffices.
Adversarial masking, force label	Tests the failure of label-preserving masking.
SAAM-EC w/o gate	Uses rule-only $\tilde{s}_q$ labels; tests interpolation.
SAAM-EC w/o abstention objective	Tests abstention training for insufficient masks.
SAAM-EC w/o extractability guard	Sets $\omega_q = \tilde{s}_q$; tests whether span visibility matters.
SAAM-EC w/o bridge units	Tests long-range evidence.
SAAM-EC w/o same-speaker units	Tests speaker-history reliance.
Full SAAM-EC	Main model.

Table 7: Controlled variants used in the SAAM-EC ablation study.

B Experimental Setup Details

B.1 Backbone Configurations

Table 8 summarizes the backbone configurations used for reproduction. SAAM-EC is inserted only into the cause-span or pair-decision training stage. The upstream emotion recognizer, candidate construction rule, and inference-time decoder of each backbone are kept unchanged for the corresponding clean and SAAM-EC runs.

For ECF, we follow each backbone’s original development-set construction: UIC holds out 20% of the official training data for development, while THOR uses its 9:1 train/dev construction. For RECCON, models are trained on RECCON-DD with an 80/10/10 train/dev/test split; RECCON-IE is never used for tuning and serves only as the cross-domain test set. Published scores are reported only as external references; comparisons involving SAAM-EC are based on the reproduced backbone under the same split and decoder.

B.2 Task-Specific Masked Supervision

SAAM-EC uses the native output space of each backbone. For extractive QA backbones, a masked positive example keeps the original span target only when the remaining input is both sufficient and extractable:

$$\omega_q = \kappa_q \tilde{s}_q, \quad \kappa_q = \mathbb{I} \left[\sum_{t \in c_q} m_{q,t} = 0 \right],$$

where c_q is the gold cause span and $m_{q,t} = 1$ means that token t is masked. If $\omega_q = 1$, the masked example is trained toward the original start and end positions. If $\omega_q = 0$, it is trained as NO_ANSWER. This target means that the original

annotated span is no longer extractable from the masked input; it does not mean that the underlying causal event is false.

THOR-ECAC is treated differently because its cause module is sequence-to-sequence classification over emotion labels plus NEUTRAL, followed by rule-based span correction. For sufficient masked inputs, the original THOR-cause-rr target is preserved. For destructive masked inputs, the target is changed to NEUTRAL. No start/end span head is added to THOR.

For PRG-MoE and CENTER, the output is an emotion-cause utterance pair. We therefore use an utterance-level extractability guard:

$$\kappa_q^{\text{pair}} = \mathbb{I} \left[\sum_{t \in u_j} (1 - m_{q,t}) > 0 \right],$$

$$\omega_q^{\text{pair}} = \kappa_q^{\text{pair}} \tilde{s}_q.$$

where u_j is the gold cause utterance. If $\omega_q^{\text{pair}} = 1$, the masked example keeps the positive pair label. Otherwise, it is trained as NO_PAIR, i.e., the existing negative pair label. Thus SAAM-EC does not introduce a new RECCON label space.

B.3 SAAM-EC Hyperparameters

Table 9 gives the default SAAM-EC hyperparameters. The selected setting is intentionally conservative: $K = 2$ removes two evidence units while avoiding the regime where most positive examples become destructive; $\lambda_{\text{mask}} = 0.3$ keeps clean supervision dominant; and $\beta_{\text{gate}} = 0.1$ prevents weak sufficiency labels from overwhelming the main extraction objective.

Thresholds are selected on the corresponding development set and then fixed for test evaluation.

Backbone	Reproduction configuration	SAAM-EC attachment
UIC Two-Step QA	RoBERTa-base emotion classifier with weighted cross-entropy, AdamW, learning rate 5×10^{-5} , weight decay 0.01, 500 warm-up steps, and 20 epochs. Cause extraction uses SpanBERT-base-cased initialized from the SQuAD 2.0 fine-tuned checkpoint, maximum sequence length 400, batch size 12, learning rate 2×10^{-5} , and 5 epochs.	Applied to the SpanBERT cause-span QA stage. The RoBERTa emotion classifier is kept unchanged.
THOR-ECAC	Flan-T5-base with THOR-state pre-tuning and THOR-cause-rr fine-tuning. We keep the original context window $k = 3$, 9:1 train/dev construction from the training file, AdamW optimizer, learning rate 2×10^{-4} , batch size 32, temperature 1.0, and rule-based span correction.	Applied to THOR-cause-rr input masking. Since THOR predicts an emotion/no-cause label before rule-based span correction, destructive masked examples are mapped to the existing NEUTRAL target rather than to a new extractive head.
PRG-MoE	bert-base-cased utterance encoder, Adam optimizer, 40 epochs, exponential learning-rate decay 0.05, learning rate 5×10^{-5} , a batch size of 5 dialogues, dropout 0.5, window size 3, and loss weights $\lambda_{\text{emo}} = 0.2$, $\lambda_{\text{pair}} = 0.8$.	Applied to pair classification. The original window-constrained pair construction is kept.
CENTER	bert-base-cased with hidden size 768. We keep Adam with learning rate 10^{-3} , 40 epochs, center-event influence radius $\beta_{\text{center}} = 3$, pair-reasoning radius $\delta = 2$, center-event annotation threshold $\alpha = 2$, and task weights $\gamma_e = \gamma_c = \gamma_{ce} = \gamma_{\text{pair}} = 1$.	Applied to final pair classification after the original CED, CEA, and CER modules. No center-event graph component is changed.

Table 8: Backbone reproduction settings. SAAM-EC changes the training objective and masked-input construction, but not the inference-time architecture.

Hyperparameter	Value	Scope
Semantic mask budget K	2	all SAAM-EC runs
Masked-loss weight λ_{mask}	0.3	all SAAM-EC runs
Gate-loss weight β_{gate}	0.1	all SAAM-EC runs
EMA decay ρ	0.999	all SAAM-EC runs
Clean warm-up N_{warm}	larger of 1 epoch and 10% of updates	before masking
SoftTopK temperature τ_{ST}	1.0 $\rightarrow$ 0.5	after warm-up
Masker/gate hidden size	256	all SAAM-EC runs
Unit-type embedding size	32	all SAAM-EC runs
Masker/gate dropout	0.1	all SAAM-EC runs
Masker/gate optimizer	AdamW, learning rate 10^{-4} , weight decay 0.01	all SAAM-EC runs
QA negative ratio r	3 negatives per positive	candidate-QA only
Maximum decoded span length L_{max}	30 subword tokens	extractive QA only
Emotion threshold τ_e	dev-tuned	ECF emotion filtering
Answerability threshold τ_z	dev-tuned	ECF QA filtering
Pair threshold τ_p	dev-tuned	RECCON pair filtering
Number of random seeds	3	reported means
GPU	single NVIDIA A100	all experiments

Table 9: Default SAAM-EC hyperparameters. We use λ_{mask} and β_{gate} to avoid confusion with backbone-specific symbols such as CENTER’s β_{center} .

For ECF, τ_e and τ_z are tuned by development w-PF1. For RECCON, τ_p is tuned by RECCON-DD development F1 and reused for both RECCON-DD and RECCON-IE test evaluation.

B.4 Mask Replacement and Tuning Policy

The notation [EVI_MASK] denotes neutral evidence replacement. In implementation, it is mapped to the native mask or sentinel token of the backbone tokenizer: [MASK] for BERT/SpanBERT-style encoders and <extra_id_0> for Flan-T5. We replace lexical tokens rather than deleting them, so token alignment and span boundaries remain valid for extractive QA. Speaker names, utterance IDs, separators, query tokens, emotion-utterance markers, candidate-cause markers, and other structural tokens are never masked.

Only the two most important method-level hyperparameters are tuned in the main sensitivity study:

$$K \in \{1, 2, 3, 4\},$$

$$\lambda_{\text{mask}} \in \{0.1, 0.3, 0.5, 1.0\}.$$

This grid is run on ECF with UIC Two-Step QA and selected by development w-PF1. After selection, the same K and λ_{mask} are used for THOR-ECAC,

PRG-MoE, and CENTER. All other values in Table 9 are fixed unless an ablation explicitly changes them.

During the backbone/gate update, the sampled hard mask is treated as fixed. During the masker update, backbone and gate parameters are frozen. The main model uses no entropy regularizer, budget penalty, degeneracy penalty, LLM verifier, or adversarial parameter perturbation. This keeps SAAM-EC close to adversarial data augmentation while isolating the effect of sufficiency-aware target reassignment.

C Additional Ablation Details

This appendix reports compact derived statistics for the ablation studies in Section 4.2. The full metric tables are given in Tables 3 and 4; here we focus on primary-metric deltas, which make the contribution of each design choice easier to compare. ECF scores are reported on the percentage scale, and RECCON scores are pair-level percentages.

Table 10 shows that force-label masking provides only small gains over clean training, even when the masked units are selected by teacher importance or by the adversarial selector. In contrast, RandomMask-SAAM gives substantially larger improvements, indicating that sufficiency-aware target reassignment is more important than masking alone. Full SAAM-EC further improves over RandomMask-SAAM on all three primary metrics, showing that adversarial evidence selection is useful when it is coupled with sufficiency-aware supervision.

The largest drops in Table 11 come from removing the extractability guard and the abstention-objective branch. This matches the motivation of SAAM-EC: after masking removes the annotated span or the gold cause utterance, preserving the original target creates false supervision. The EMA-teacher features have a smaller but consistent effect, suggesting that they mainly stabilize sufficiency estimation. Discourse units also matter: removing bridge and same-speaker evidence hurts RECCON-IE most, which is consistent with the stronger need for conversational context under cross-domain shift.

Variant	Δ ECF w-PF1	Δ DD F1	Δ IE F1
RandomMask-Force	+0.28	+0.16	+0.61
ImportantMask-Force	+0.08	+0.14	+1.01
AdvMask-Force	+0.03	+0.19	+0.96
RandomMask-SAAM	+1.37	+0.75	+2.33
SAAM-EC	+1.85	+1.25	+3.16

Table 10: Primary-metric gains over the clean backbone in the masking ablation. ECF gains are absolute w-PF1 points; DD and IE denote RECCON-DD and RECCON-IE pair F1.

Variant	Δ ECF w-PF1	Δ DD F1	Δ IE F1
w/o sufficiency gate	-0.40	-0.46	-0.87
w/o abstention objective	-0.91	-0.82	-0.93
w/o extractability guard	-1.03	-1.00	-0.99
w/o EMA features	-0.18	-0.25	-0.49
w/o bridge units	-0.45	-0.61	-0.99
w/o same-speaker units	-0.31	-0.43	-0.64
w/o discourse units	-0.62	-0.82	-1.29

Table 11: Primary-metric drops from full SAAM-EC in the component ablation. Negative values indicate degradation relative to the full model.

D Detailed Evidence-intervention Diagnostics

This appendix gives the exact diagnostic tables behind Figure 4. The analysis uses ECF 2.0 with UIC Two-Step QA and compares the clean backbone, AdvMask-Force, and SAAM-EC. The diagnostic is evaluation-only: it does not change the trained models.

D.1 Diagnostic Set Construction

Quantity	Count
Test conversations	341
Positive emotion-cause pair instances	2,256
Unique positive query groups	2,241
Multi-span query groups	15
Positive instances in multi-span groups	30

Table 12: Positive-query pool used for ECF evidence-intervention diagnostics.

For each positive query, the sufficient intervention masks shortcut evidence while keeping the gold cause span visible. The destructive interventions remove either the full gold span or the full candidate cause utterance. Under destructive masking, NO_ANSWER is the correct extractive target because the original annotated span is no longer recoverable from the input.

Intervention	Masking operator, target, and purpose	n
Sufficient mask	Shortcut evidence masked while the gold span remains visible; target: original span; purpose: span recovery.	2,256
Destructive mask	Full gold cause span removed; target: NO_ANSWER; purpose: abstention.	2,256
Destructive mask	Full candidate cause utterance removed; target: NO_ANSWER; purpose: abstention.	2,256
Destructive pooled	Span removal plus candidate-utterance removal; target: NO_ANSWER; purpose: hallucination analysis.	4,512

Table 13: Intervention instances in the ECF diagnostic.

Sufficient-mask subtype	# instances	Share
Explicit emotion cue masked	628	27.84%
Local causal template masked	599	26.55%
Nearby non-gold distractor masked	1,029	45.61%
Total	2,256	100.00%

Table 14: Subtype counts for sufficient masks.

D.2 Recovery Under Sufficient Masking

Model	n	Recovery / answerable / conditional PF1	SD	Δ C / A
Clean Backbone	2,256	21.68 / 84.13 / 25.77	0.28	0.00 / -
AdvMask-Force	2,256	22.89 / 86.22 / 26.54	0.26	+1.21 / 0.00
SAAM-EC	2,256	25.29 / 90.06 / 28.08	0.16	+3.61 / +2.40

Table 15: Recovery under sufficient masking on ECF. Metric triples report recovery w-PF1, answerable rate, and conditional span PF1, respectively. Deltas are relative to Clean Backbone and AdvMask-Force; scores are percentages except counts.

Table 15 shows that SAAM-EC improves recovery when shortcut evidence is masked but the gold span remains visible. The gain comes from both answerability and span localization: SAAM-EC predicts more recoverable examples as answerable and also obtains higher conditional span PF1 among answerable predictions.

The subtype results show that the recovery gain is not confined to one shortcut family. SAAM-EC improves after masking explicit emotion cues, local causal templates, and nearby non-gold distractors.

Mask subtype	n	Answerable / conditional / recovery
Clean Backbone		
Explicit emotion cue	628	84.60 / 26.06 / 22.05
Local causal template	599	81.80 / 25.62 / 20.96
Nearby non-gold distractor	1,029	85.20 / 25.68 / 21.88
AdvMask-Force		
Explicit emotion cue	628	87.30 / 26.63 / 23.25
Local causal template	599	84.10 / 26.63 / 22.40
Nearby non-gold distractor	1,029	86.80 / 26.44 / 22.95
SAAM-EC		
Explicit emotion cue	628	91.20 / 28.06 / 25.59
Local causal template	599	88.80 / 27.97 / 24.84
Nearby non-gold distractor	1,029	90.10 / 28.16 / 25.37

Table 16: Recovery under sufficient masks by shortcut subtype. Metric triples report answerable rate, conditional span PF1, and recovery w-PF1, respectively; all scores are percentages.

D.3 Hallucination Under Destructive Masking

Model	n	Hallucination / SD / NO_ANSWER recall	Median answerable	Δ Clean
Clean Backbone	4,512	60.68 / 0.89 / 39.32	2,738 / 4,512	0.00
AdvMask-Force	4,512	67.18 / 0.87 / 32.82	3,031 / 4,512	+6.50
SAAM-EC	4,512	27.24 / 0.97 / 72.76	1,229 / 4,512	-33.44

Table 17: Hallucination under destructive masking. Lower hallucination and higher NO_ANSWER recall are better. Rates are percentages.

AdvMask-Force hallucinates more often than the clean backbone under destructive masking, which is consistent with its label-preserving training objective. By contrast, SAAM-EC substantially increases NO_ANSWER recall and reduces unsupported answerable predictions.

SAAM-EC reduces hallucinations at residual candidate locations relative to AdvMask-Force. The remaining errors are still often associated with nearby non-gold distractors, indicating that distractor suppression remains an important failure mode.

Destructive operator	n	Answerable count / hallucination / mean p
Clean Backbone		
Full gold span removed	2,256	1,478 / 65.51 / 64.10
Full candidate utterance removed	2,256	1,260 / 55.85 / 58.60
AdvMask-Force		
Full gold span removed	2,256	1,619 / 71.76 / 70.20
Full candidate utterance removed	2,256	1,412 / 62.59 / 64.10
SAAM-EC		
Full gold span removed	2,256	696 / 30.85 / 35.20
Full candidate utterance removed	2,256	533 / 23.63 / 28.20

Table 18: Hallucination by destructive operator. Result triples report answerable count, hallucination rate, and mean answerability probability; the latter two are percentages.

Hallucinated source type	Count	Share
Clean Backbone		
Residual candidate-location span	1,179	43.1%
Nearest non-gold distractor span	883	32.3%
Emotion-utterance cue span	411	15.0%
Other / unstable span	265	9.7%
AdvMask-Force		
Residual candidate-location span	1,574	51.9%
Nearest non-gold distractor span	822	27.1%
Emotion-utterance cue span	397	13.1%
Other / unstable span	238	7.9%
SAAM-EC		
Residual candidate-location span	439	35.7%
Nearest non-gold distractor span	444	36.1%
Emotion-utterance cue span	216	17.6%
Other / unstable span	130	10.6%

Table 19: Source breakdown of hallucinated answers under destructive masks.

D.4 Hard-case Bucket Results

Bucket	Definition	n	Overlap note
Same-utterance	$i = j$	1,132	distance partition
Adjacent	$ i - j = 1$	675	distance partition
Distant	$ i - j \geq 2$	449	distance partition
Cross-speaker	$s_i \neq s_j$	743	overlaps with distance buckets

Table 20: Hard-case bucket definitions for ECF diagnostics.

The bucket results show a clear difficulty pattern. SAAM-EC gives smaller gains on same-utterance cases, where local evidence is often sufficient, and larger gains on distant and cross-speaker cases, where the model must rely on broader conversa-

tional evidence rather than proximity or speaker shortcuts.

Bucket	n	w-PF1 C / A / S	Δ S-C / A-C	95% CI
Same-utterance	1,132	27.93 / 27.61 / 28.86	+0.93 / -0.32	[+0.24, +1.65]
Adjacent	675	24.31 / 24.55 / 26.17	+1.86 / +0.24	[+0.94, +2.71]
Distant, $ i - j \geq 2$	449	15.48 / 16.41 / 19.74	+4.26 / +0.93	[+3.06, +5.51]
Cross-speaker	743	20.76 / 20.92 / 23.87	+3.11 / +0.16	[+2.09, +4.19]

Table 21: Main hard-case bucket gains. C/A/S denotes Clean Backbone, AdvMask-Force, and SAAM-EC; scores and intervals are percentage points.

Bucket	n	Clean S / P	AdvMask S / P	SAAM-EC S / P	Δ SAAM S / P
Same-utterance	1,132	21.51 / 27.93	21.02 / 27.61	22.09 / 28.86	+0.58 / +0.93
Adjacent	675	18.38 / 24.31	18.05 / 24.55	19.37 / 26.17	+0.99 / +1.86
Distant, $ i - j \geq 2$	449	10.48 / 15.48	11.25 / 16.41	13.89 / 19.74	+3.41 / +4.26
Cross-speaker	743	14.56 / 20.76	14.68 / 20.92	17.69 / 23.87	+3.13 / +3.11

Table 22: Hard-case bucket results with strict (S) and proportional (P) F1. Each cell reports w-SF1 / w-PF1; scores are percentages.

Bucket	n	w-PF1 C / A / S	Δ S-C
Distance = 0	1,132	27.93 / 27.61 / 28.86	+0.93
Distance = 1	675	24.31 / 24.55 / 26.17	+1.86
Distance = 2	243	17.58 / 17.12 / 20.53	+2.95
Distance ≥ 3	206	13.03 / 15.57 / 18.80	+5.77
Same-speaker	1,513	26.22 / 26.18 / 27.45	+1.23
Cross-speaker	743	20.76 / 20.92 / 23.87	+3.11

Table 23: Expanded bucket results. C/A/S denotes Clean Backbone, AdvMask-Force, and SAAM-EC; scores are percentages.

E Additional Sensitivity Results

Table 24 provides the full sensitivity grid over the semantic mask budget K and masked-loss weight λ_{mask} . The clean references are ECF w-SF1/w-PF1 = 18.39/24.42, RECCON-DD P/R/F1 = 58.95/55.67/57.26, and RECCON-IE P/R/F1 = 51.95/20.02/28.90. ECF uses UIC Two-Step QA, and RECCON-DD/IE use PRG-MoE.

K	λ_{mask}	ECF w-SF1 / w-PF1	RECCON-DD P / R / F1	RECCON-IE P / R / F1
1	0.1	19.02 / 25.57	59.31 / 56.59 / 57.92	52.10 / 21.43 / 30.37
1	0.3	19.30 / 25.91	59.53 / 56.98 / 58.23	52.34 / 22.30 / 31.27
1	0.5	19.17 / 25.78	59.38 / 56.84 / 58.08	51.88 / 22.68 / 31.56
1	1.0	18.74 / 25.19	59.12 / 56.25 / 57.65	50.86 / 22.72 / 31.41
2	0.1	19.36 / 26.04	59.49 / 57.06 / 58.25	52.42 / 22.47 / 31.46
2	0.3	19.62 / 26.27	59.71 / 57.36 / 58.51	52.47 / 23.08 / 32.06
2	0.5	19.50 / 26.13	59.60 / 57.19 / 58.37	52.06 / 23.00 / 31.90
2	1.0	18.95 / 25.47	59.22 / 56.46 / 57.81	50.94 / 22.82 / 31.52
3	0.1	19.19 / 25.85	59.37 / 56.79 / 58.05	52.06 / 22.16 / 31.09
3	0.3	19.44 / 26.08	59.56 / 57.09 / 58.30	52.21 / 22.72 / 31.66
3	0.5	19.26 / 25.92	59.43 / 56.91 / 58.14	51.69 / 22.85 / 31.69
3	1.0	18.69 / 25.16	58.98 / 56.07 / 57.49	50.32 / 22.43 / 31.03
4	0.1	18.87 / 25.44	59.11 / 56.39 / 57.72	51.53 / 21.80 / 30.64
4	0.3	19.06 / 25.61	59.26 / 56.55 / 57.87	51.65 / 22.29 / 31.14
4	0.5	18.81 / 25.32	59.00 / 56.10 / 57.51	50.62 / 22.13 / 30.80
4	1.0	18.42 / 24.69	58.63 / 55.48 / 57.01	49.38 / 21.56 / 30.02

Table 24: Full hyperparameter sensitivity grid. Slash-separated scores follow the metric order shown in each column header; all scores are percentages. The highlighted default is $K = 2, \lambda_{\text{mask}} = 0.3$.

The best cell improves over clean training by +1.85 ECF w-PF1, +1.25 RECCON-DD F1, and +3.16 RECCON-IE F1. Moderate masking is stable, especially for $K \in \{1, 2, 3\}$ and $\lambda_{\text{mask}} \in \{0.1, 0.3, 0.5\}$. Performance declines when masking becomes too aggressive, such as $K = 4$ or $\lambda_{\text{mask}} = 1.0$, supporting the design choice that evidence intervention should disrupt shortcut reliance without making most positive examples insufficient.

F Qualitative Case Details

This appendix gives the numerical details for the qualitative case in Figure 5. The example is from the ECF 2.0 / TECPE test set and targets the emotion instance U12_anger.

F.1 Case Metadata and Dialogue

Field	Value
Dataset	ECF 2.0 / TECPE test set
Conversation ID	1727
Target emotion	U12_anger
Emotion speaker	Monica
Gold cause utterance	U7
Gold cause speaker	Chandler
Case type	distant + cross-speaker + bridge-mediated
Distance	$ 12 - 7 = 5$
Gold causes for target	1

Table 25: Metadata for the qualitative evidence-intervention case.

Turn	Speaker	Role	Text
U7	Chandler	gold cause	You know? So [GOLD] I was thinking we could get one of those signs and hang it over our bed. [GOLD]
U8	Chandler	bridge context	Because that is you and I together.
U9	Chandler	local shortcut	“Merge.”
U10	Monica	nearby same-speaker distractor	Oh, my God, I love that!
U11	Chandler	local prompt	Really?
U12	Monica	target emotion: anger	“No!”

Table 26: Intervention-relevant dialogue excerpt. The model receives the full dialogue; only the relevant turns are shown here.

The gold cause is the distant cross-speaker U7 span. The example is difficult because U8 provides bridge context, U9 supplies a local shortcut phrase, and U10 is a nearby same-speaker affective distractor.

F.2 Intervention States and Model Outputs

Intervention	Mask state, visibility, and target / purpose	Ratio
Clean	No evidence masked; gold span visible: yes; original test input; target: original span.	0.000
Sufficient mask	U9 “Merge.” and U10 “I love that!” masked; gold span visible: yes; U7 remains extractable; target: original span.	0.123
Destructive mask	Full U7 candidate utterance masked; gold span visible: no; extractive gold evidence removed; target: NO_ANSWER.	0.198

Table 27: Intervention states for the qualitative case. Mask ratio is computed over dialogue lexical tokens.

Model	p_{top} / candidate	Predicted output; outcome
Clean		
Clean	0.735 / U10	target: original span I love that!; <i>wrong: nearby same-speaker distractor</i>
AdvMask-Force	0.798 / U8	that is you and I together.; <i>wrong: bridge span, non-gold for U12_anger</i>
SAAM-EC	0.873 / U7	I was thinking we could get one of those signs and hang it over our bed; <i>correct</i>
Sufficient mask		
Clean	0.668 / U8	target: original span that is you and I together.; <i>wrong: bridge span, non-gold for U12_anger</i>
AdvMask-Force	0.816 / U7	get one of those signs and hang it over our bed; <i>partial recovery</i>
SAAM-EC	0.903 / U7	I was thinking we could get one of those signs and hang it over our bed; <i>correct</i>
Destructive mask		
Clean	0.579 / U8	target: NO_ANSWER Because that is you and I together.; <i>hallucinated answer</i>
AdvMask-Force	0.681 / U7	[EVI_MASK]; <i>hallucinated masked-location answer</i>
SAAM-EC	0.159 / none	NO_ANSWER; <i>correct abstention</i>

Table 28: Final model outputs for the qualitative intervention case. p_{top} is the maximum answerability score over candidate cause utterances for U12_anger; the visualization threshold is $\tau_z = 0.50$.

On this case, the clean backbone follows nearby or bridge-like shortcuts (U10/U8), while AdvMask-Force recovers U7 under the sufficient mask but still answers after its evidence is removed. SAAM-EC separates the regimes, recovering U7 when visible and outputting NO_ANSWER after destructive masking.

F.3 Candidate-score Values

Model	Candidate scores U7 / U8 / U10 / U12	Decision
Clean		
Clean Backbone	0.612 / 0.671 / 0.735 / 0.704	U10
AdvMask-Force	0.775 / 0.798 / 0.736 / 0.688	U8
SAAM-EC	0.873 / 0.594 / 0.512 / 0.408	U7
Sufficient mask		
Clean Backbone	0.541 / 0.668 / 0.162 / 0.636	U8
AdvMask-Force	0.816 / 0.743 / 0.151 / 0.591	U7
SAAM-EC	0.903 / 0.521 / 0.103 / 0.345	U7
Destructive mask		
Clean Backbone	0.438 / 0.579 / 0.551 / 0.533	U8
AdvMask-Force	0.681 / 0.641 / 0.598 / 0.562	U7
SAAM-EC	0.118 / 0.159 / 0.142 / 0.131	NO_ANSWER

Table 29: Candidate-score values for the qualitative case, ordered U7/U8/U10/U12.