They Are Not the Same: Direct-Cause Extraction Is Not Evidence Grounded Emotion Explanation

Anonymous ACL submission

Abstract

Emotion-Cause Pair Extraction (ECPE) was introduced to explain why an emotion occurs, but this goal is now often reduced to binary pair/non-pair prediction. This proxy is useful for direct-cause extraction, yet easy to over-read as evidence grounded emotion explanation. We show that this interpretation is only partially valid. In IEMO-MECP, 90.9% of original positives remain `emo-cause` and 95.0% of original negatives remain `non-pair`, confirming that the binary ECPE task is largely preserved. The problem is that direct triggers alone do not constitute a grounded explanation of emotion. `emo-context`, an utterance that helps interpret a target emotion without directly causing it, appears on both sides of the original boundary and is enriched near binary uncertainty, showing that the binary boundary has no stable place for such discourse evidence. Across evaluated ECPE models, direct triggers are recovered more reliably than contextual support. Under shortcut pressure, this imbalance becomes consequential. Binary-trained models assign higher pair scores to nearby lexically similar `non-pair` candidates than to harder `emo-cause` and `emo-context` pairs that provide stronger explanatory evidence. In other words, pair scores can reward convenient attributions over explanatory evidence. Thus, high binary ECPE performance reflects direct-trigger identification, not emotion explanation. Code and dataset will be released.

1 Introduction

Emotion-Cause Pair Extraction (ECPE) was introduced to explain why an emotion occurs by recovering emotion-cause pairs from text (Xia and Ding, 2019). The task has since moved from documents to conversations, where causes may depend on speaker turns, interaction history, and dialogue structure (Poria et al., 2021; Li et al., 2023; Jeong and Bak, 2023), and further to multimodal conversations that combine textual, acoustic, and visual cues (Wang et al., 2023, 2024b; Li et al., 2025).

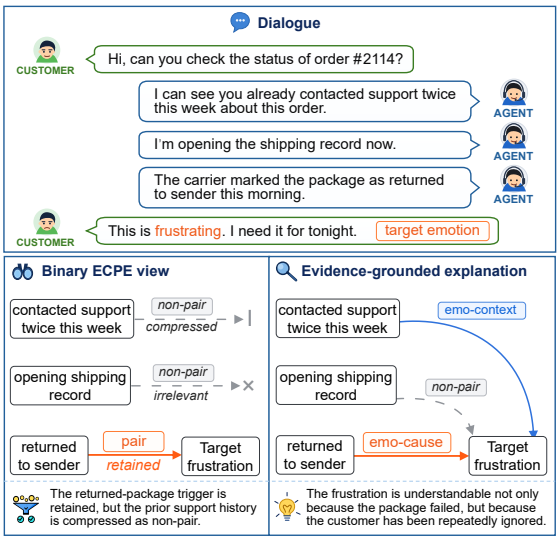

Figure 1: Boundary-compression example. Binary ECPE can retain the direct trigger while compressing discourse evidence that makes the emotion interpretable. The validity risk is not that the trigger is missed, but that trigger extraction is treated as an emotion explanation.

Despite these advances, progress on this explanation-oriented task is commonly evaluated through a narrower proxy, namely binary pair/non-pair prediction. This proxy is useful for direct-cause extraction, but it changes the measurement target. A model can succeed at deciding whether an utterance should be counted as a cause pair while still failing to preserve the discourse evidence that makes the emotion understandable in context. The resulting risk is not simply lower performance on hard cases. It is an ***evaluation validity failure***, where pair scores may certify trigger extraction while being interpreted as evidence grounded emotion explanation. Li et al. (2018); Yan et al. (2021); Zhao et al. (2023) showed that context, position, and commonsense evidence affect emotion-cause reasoning. Our concern is whether the binary pair/non-pair label space can represent discourse evidence that is explanatory but not a direct trigger.

Figure 1 illustrates this gap. The prior contact utterance is not a weaker form of the returned package cause. It does not trigger the final frustration, but changes how that frustration is understood. The target utterance reads as part of an ongoing unresolved service problem. This is the `emo-context` role, contextual support that helps interpret a target emotion without directly causing it. The question is therefore not whether binary ECPE is useful, but what its pair scores can justify. A high pair score may support a direct trigger decision, but should not be read as evidence that the model has recovered an evidence grounded explanation of the emotion. At the label-space level, this is a projection problem, where binary ECPE must place direct causes, contextual support, and unsupported candidates into only two labels. As a result, contextual support is either over-causalized as *pair* or discarded as *non-pair*, and neither choice preserves the full explanation structure.

We study this problem with IEMO-MECP as a diagnostic audit of the inherited ECPE candidate space, not as a replacement for existing benchmarks. The audit keeps the original target-candidate pairs and source-binary labels, then assigns each pair one of three roles, `emo-cause`, `emo-context`, or `non-pair`. This gives a binary view and a role-aware view of the same candidate set, aligned pair by pair. This alignment matters because disagreements between the two views can be analyzed as evidence about what the binary boundary compresses, not as artifacts of different candidate construction. Our goal is to identify where binary pair scores remain valid, where their emotion explanation claims end, and why this boundary matters for evaluating emotion explanation systems.

Our contributions are threefold.

- We frame binary ECPE scoring as an evaluation validity problem, showing that pair scores support direct-cause decisions but not evidence grounded emotion explanation.
- We introduce IEMO-MECP as a diagnostic audit of the inherited ECPE candidate space, where `emo-context` marks non-triggering discourse evidence compressed by the binary boundary.
- We evaluate ECPE models with role-wise diagnostics and shortcut stress tests, showing that `emo-context` often collapses into `non-pair` and that local shortcuts can outrank stronger explanatory evidence.

2 Evaluation-Validity Framing

2.1 Pair Scores and Explanation Claims

We formalize the gap between binary pair decisions and claims of emotion explanation. Let $D = \{u_1, \cdots, u_n\}$ be a dialogue and let $\mathcal{E}(D)$ denote its non-neutral target emotion turns. Following conversational ECPE, the candidate space is lower triangular:

$$\mathcal{P}(D) = \{(i, j) : i \in \mathcal{E}(D),\ 1 \leq j \leq i\}, \quad (1)$$

where u_i is the target emotion utterance and u_j is either the target utterance itself or an earlier candidate. Standard binary ECPE assigns $y_{ij}^{\text{bin}} \in \{0, 1\}$, indicating whether u_j is treated as a cause of the emotion in u_i. We use pair score to denote the confidence assigned to $y_{ij}^{\text{bin}} = 1$. This score supports a binary pair decision, namely whether u_j should be counted as a cause of the emotion in u_i.

The decision is appropriate when the evaluation target is direct-cause extraction. The validity problem arises when the same score is used as evidence that a model explains why the target emotion occurred. Contextual cues can shape how a target emotion is interpreted without being the direct trigger. Pair-level ECPE evaluation, therefore, needs a distinction between direct triggers and non-triggering discourse evidence.

2.2 Defining Pair Roles

We assign each candidate pair a role $y_{ij}^{\text{role}} \in \mathcal{R}$, where $\mathcal{R} = \{c, x, n\}$, such that c, x, and n correspond to `emo-cause`, `emo-context`, and `non-pair`. These roles are not confidence levels. They describe different evidential relations between a candidate utterance and a target emotion. Let D_{ij} denote observable direct-cause evidence, where the candidate supplies the direct trigger, event, or appraisal for the target emotion. Let S_{ij} denote observable discourse evidence that helps interpret the target emotion without directly triggering it. The intended role hierarchy is

$$y_{ij}^{\text{role}} = \begin{cases} c, & D_{ij}, \\ x, & \neg D_{ij} \wedge S_{ij}, \\ n, & \text{otherwise.} \end{cases} \quad (2)$$

If a candidate both supplies background and directly grounds the target emotion, it is labeled `emo-cause`. The `emo-context` role is reserved for evidence that supports interpretation without serving as the direct trigger.

Table 1: Pair-role criteria for the diagnostic audit.

Role	Positive criterion	Exclusion criterion	Frequent boundary error
emo-cause	Direct trigger, event, or appraisal for the target emotion. Self-pairs are valid when the target utterance contains this evidence.	Not for background or contextual support alone.	Over-causalizing background evidence.
emo-context	Emotionally relevant but non-triggering; removal weakens target-emotion interpretation.	Not for weak topical similarity, proximity, or generic prior dialogue.	Treating context as weak cause or soft non-pair.
non-pair	Insufficient evidence for direct causality or contextual emotional relevance.	Not when the candidate supplies background, conflict setup, relationship state, or situational evidence.	Discarding useful non-triggering discourse evidence.

Table 1 turns this hierarchy into audit criteria. The key boundary is emo-context. A emo-context pair must make the target emotion easier to interpret when included and harder to interpret when removed. It may involve background events, conflict setup, relationship state, discourse bridging, prior speaker state, situational constraints, or alternative plausible causes that shape appraisal. Mere proximity or topical similarity is not enough. This keeps emo-context separate from both direct causes and ordinary non-pair candidates.

Binary projection loses one distinction. The role space contains three evidential relations, $\mathcal{R} = \{\texttt{emo-cause}, \texttt{emo-context}, \texttt{non-pair}\}$, while binary ECPE observes only $y^{\text{bin}} \in \{0, 1\}$. Any binary evaluation, therefore, applies a projection $\pi : \mathcal{R} \to \{0, 1\}$. For direct-cause extraction, the natural projection satisfies $\pi(\texttt{emo-cause}) = 1$ and $\pi(\texttt{non-pair}) = 0$. The role emo-context cannot be projected into the binary pair/non-pair label space without losing either its non-causal status or its function in emotion explanation. If $\pi(\texttt{emo-context}) = 1$, contextual support is over-causalized as a pair; if $\pi(\texttt{emo-context}) = 0$, it is treated as irrelevant. Thus, binary scoring can preserve the direct-cause decision while losing the distinction between unsupported candidates and non-triggering explanatory evidence. See Appendix B for completeness.

2.3 Diagnostic Predictions

To make the validity question testable, we turn the role distinction into checks on audited labels and model scores, yielding five diagnostic predictions.

P1. Direct-cause preservation. If the role-aware view refines rather than replaces the binary task, original positives should largely remain emo-cause, and original negatives should largely remain non-pair.

P2. Boundary crossing. If the binary boundary compresses contextual support, emo-context should appear among both original positives and original negatives.

P3. Uncertainty enrichment. If emo-context is compressed by the binary boundary, it should be enriched near binary uncertainty rather than binary-score extremes.

P4. Context-channel weakness. If the binary objective lacks a context channel, gold emo-context pairs should score closer to non-pair than to emo-cause.

P5. Shortcut-evidence conflict. If pair scores partly reflect shortcut-compatible cues, unsupported local candidates may outrank evidence-supported but structurally harder pairs.

3 Role Audit and Boundary Diagnostics

3.1 Role-Audited Candidate Space

IEMO-MECP is a role-audited extension of ConvECPE (Li et al., 2023), an IEMOCAP-based conversational ECPE dataset (Busso et al., 2008). IEMOCAP provides the underlying multimodal dialogues and emotion annotations, while ConvECPE adds emotion-cause annotations. We keep the inherited target candidate pairs and source-binary labels, and assign each candidate pair the three-role label defined in Section 2.

Role auditing follows the controlled design used by the diagnostics. The inherited source-binary labels are retained as metadata, while role labels are assigned under the criteria in Table 1. An LLM provides initial role proposals, which are then reviewed for dialogue-level consistency, refined in high-risk boundary regions, and checked at the final file level. High-risk boundary transitions are additionally reviewed by human auditors. A blind re-annotation audit on 330 boundary-focused judgments yields 0.821 pairwise agreement, mean Cohen's κ of 0.732, and Krippendorff's α of 0.732; full provenance, judging, and draft-to-final diagnostics are in Appendices A.1-A.4.

The audited candidate space contains 86,075 pairs, with 8,635 emo-cause, 4,346 emo-context, and 73,094 non-pair pairs. Full split statistics are reported in Appendix C.1. The labeling prompt and refinement rules are reported in Appendix F.

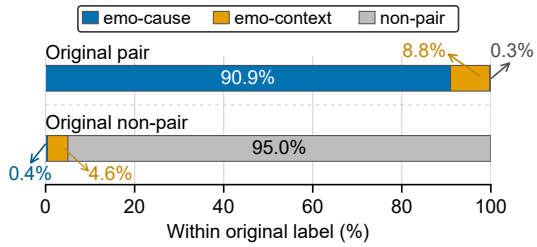

Figure 2: Source-binary transitions. The binary task is largely preserved, but `emo-context` crosses both sides of the inherited pair/non-pair boundary.

Table 2: Checks showing that `emo-context` is not a weak cause, annotation artifact, or distance heuristic.

Ruled-out explanation	Audit evidence
`emo-context` is only missing positives or weak `emo-cause`	`emo-context` appears in both original positives (8.8%) and original negatives (4.6%).
`emo-context` is an annotation artifact	All 4,346 `emo-context` pairs have explicit non-default provenance.
`emo-context` is just a distance heuristic	68.4% occur at distance ≥ 3, but 1,375 occur at shorter distances.

3.2 Source-Binary Transitions

Figure 2 compares source-binary and final pair-role labels over the same candidate space. For P1, the large preserved blocks show that the audit largely retains the original direct-cause abstraction with 90.9% of original positives remain `emo-cause` and 95.0% of original negatives remain `non-pair`. Thus, the boundary problem is not an artifact of replacing ECPE with a different task.

For P2, the smaller `emo-context` blocks carry the validity signal. Among original positives, `emo-context` marks interpretive support that was over-counted as causal. Among original negatives, it marks interpretive support that was discarded as irrelevant. This two-sided movement rules out a simple missing-positive account and also rules out treating `emo-context` as a weak `emo-cause` label. The inherited binary boundary preserves many direct-cause decisions, but it cannot keep causal status and explanatory relevance separate.

Table 2 rules out three simpler explanations for this boundary-crossing pattern. These checks show that `emo-context` is neither a weak cause, nor a default annotation artifact, nor a distance-defined label. The pattern is therefore better explained as compression of explanatory relevance by the inherited binary boundary. Full source-binary and self-pair diagnostics are in Appendix A.5.

3.3 Boundary Uncertainty

Figure 3 tests P3 using a binary-source RWC-Fusion model with RoBERTa, WavLM, and CLIP

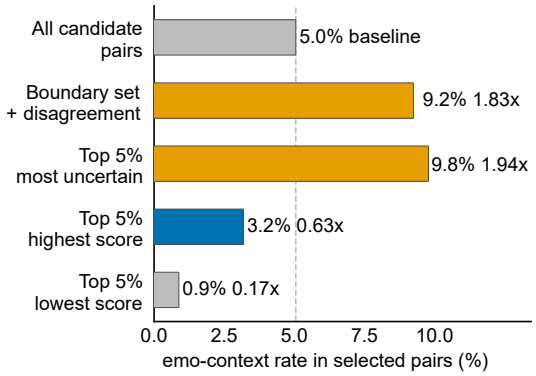

Figure 3: `emo-context` is enriched near binary uncertainty rather than at confident score extremes.

features over all 86,075 scored pairs. `emo-context` appears most often where the binary model is least certain. Its overall rate is 5.0%, but rises to 9.2% in the boundary set with source disagreement and 9.8% among the top 5% most uncertain pairs. It is much rarer at the confident extremes, appearing in only 3.2% of the top 5% highest-score pairs and 0.9% of the top 5% lowest-score pairs.

This rules out two simpler explanations. `emo-context` is not just weak `emo-cause`, because it does not concentrate among high-score pairs. It is also not ordinary `non-pair` noise, because it does not concentrate among low-score pairs. Instead, it appears where the binary model is unsure. These cases show where the binary label space absorbs discourse evidence without giving it a name.

3.4 Source-Binary Compatibility

Before using the role-aware labels to test claims of emotion explanation, we first check if they preserve the original ECPE task. Sections 3.2 and 3.3 show that `emo-context` is compressed by the inherited binary boundary. Figure 4 addresses a natural concern about whether three-role supervision gains interpretability by drifting from the source-binary objective. We remap three-role predictions to pair/non-pair labels and compare Pair F1 against models trained directly on original binary labels.

Role-aware supervision remains competitive with, and often improves over, direct binary training after remapping. The strongest results occur when the projection keeps the `emo-cause` channel separate, either by dropping predicted `emo-context` or by mapping it to `non-pair`. Mapping predicted `emo-context` into the positive pair class often hurts text-grounded settings, showing that the gains do not come from expanding the positive class with contextual support, but from sep-

Table 3: Representative role-wise model behavior over three seeds. Ctx Hit is the % of gold `emo-context` pairs predicted correctly, and Ctx→NP is the % predicted as `non-pair`. Full model-modality results and cross-family collapse diagnostics are in Appendices D.2 and E.

Model family	Mod.	Macro F1	`emo-cause` F1	`emo-context` F1	`non-pair` F1	Ctx Hit	Ctx→Cause	Ctx→NP
RoBERTa	T	$48.77_{\pm 0.39}$	$40.28_{\pm 0.52}$	$17.18_{\pm 1.47}$	$88.84_{\pm 0.29}$	18.3	8.9	72.8
WavLM	A	$40.63_{\pm 0.33}$	$28.26_{\pm 0.23}$	$3.65_{\pm 0.73}$	$89.97_{\pm 0.10}$	2.0	13.1	84.9
CLIP	V	$30.19_{\pm 1.32}$	$18.50_{\pm 0.22}$	$5.20_{\pm 2.02}$	$66.86_{\pm 3.36}$	7.8	39.5	52.8
RWC-Fusion	T+A+V	$48.41_{\pm 0.50}$	$40.89_{\pm 0.63}$	$16.94_{\pm 1.74}$	$87.41_{\pm 0.52}$	21.3	9.5	69.1
MECPE-2step	T+A	$55.60_{\pm 0.08}$	$52.98_{\pm 0.11}$	$25.14_{\pm 0.91}$	$88.67_{\pm 0.60}$	35.5	9.7	54.8
HiLo	T+A+V	$52.94_{\pm 0.61}$	$50.60_{\pm 0.69}$	$19.60_{\pm 2.06}$	$88.61_{\pm 0.99}$	25.0	14.3	60.8
M^3HG	T+A	$57.95_{\pm 0.07}$	$56.73_{\pm 0.18}$	$26.84_{\pm 0.33}$	$90.30_{\pm 0.52}$	34.0	11.0	55.0

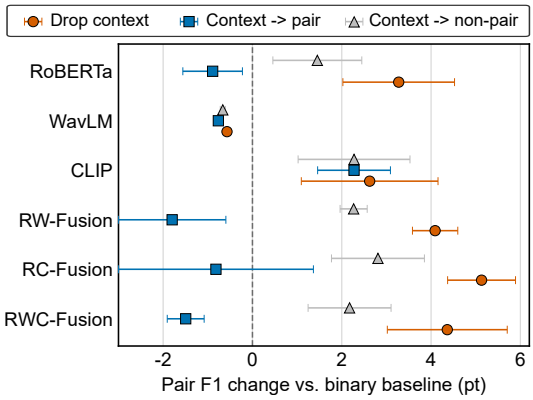

Figure 4: Pair F1 change after remapping three-role predictions to the binary ECPE task. Error bars show variation over three seeds.

arating direct-cause evidence from non-triggering discourse evidence during training.

Thus, role-aware supervision does not reject binary ECPE. It remains compatible with source-binary evaluation while separating direct-cause extraction from broader emotion-explanation claims. Absolute source-binary remapping results are in Appendix E.2.

3.5 Model Diagnostic Setup

To test whether models preserve the audited role distinction, we evaluate text (T), audio (A), and video (V) baselines using RoBERTa, WavLM, and CLIP-derived features (Liu et al., 2019; Chen et al., 2022; Radford et al., 2021), along with stronger conversational and multimodal ECPE-style systems, including HiLo, MECPE-2step, and M^3HG (Wang et al., 2023; Li et al., 2025; Liang et al., 2025). Each setting uses seeds 42, 123, and 456; tables report mean and standard deviation. As the roles are imbalanced, evaluation emphasizes macro F1, class-wise F1, and role-specific diagnostics rather than accuracy. See Appendix C for full optimization, architecture recipes and compute details.

4 Models Preserve Direct Causes More Reliably than Context

4.1 Role-Wise Results

Table 3 gives the model-side counterpart of the role audit; calibration baselines are in Appendix E.1. Three-class pair-role prediction is learnable, and stronger ECPE-style systems improve aggregate macro F1. The more important pattern is role-wise. `emo-cause` and `non-pair` are recovered much more reliably than `emo-context`, which remains the limiting role across model families.

The error direction matters. Among the representative rows, M^3HG is the strongest overall setting, with the highest Macro F1 and the highest `emo-context` F1. Yet it predicts only 34.0% of gold `emo-context` pairs correctly, while assigning 55.0% to `non-pair` and only 11.0% to `emo-cause`. If `emo-context` were merely weak `emo-cause`, missed cases should move toward `emo-cause`. Instead, most missed contexts collapse into `non-pair`, suggesting that models often treat contextual support as unsupported evidence. Appendix D.2 shows the same `emo-context`-to-`non-pair` tendency across model families.

This pattern is not explained by modality choice alone. Although aggregate performance varies across text, audio, video, and fusion settings, the central failure mode is stable. Models preserve `emo-cause` and `non-pair` distinctions more reliably than non-triggering discourse evidence. The issue is not simply adding more modality signal, but preserving `emo-context` as distinct from both direct causality and ordinary `non-pair` evidence.

4.2 Binary Scores Lack a Context Channel

Figure 5 tests P4 by locating gold `emo-context` on the score axis. Under original binary supervision, `emo-context` is much closer to `non-pair` than to `emo-cause` in mean $p(\text{pair})$. This is expected under a pair/non-pair objective, since contextual sup-

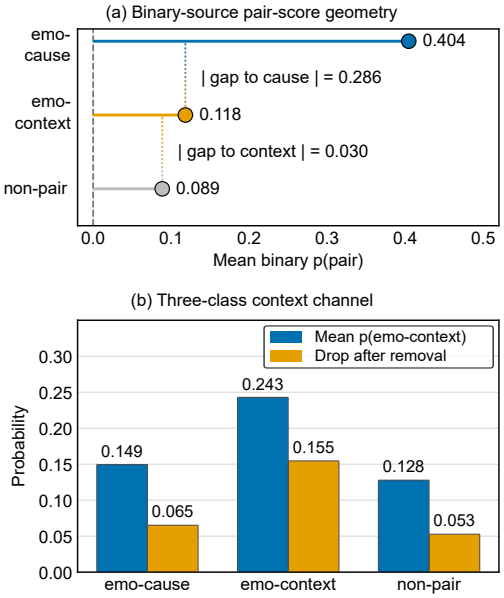

Figure 5: Context-channel diagnostics for RWC-Fusion. Binary supervision places gold `emo-context` near `non-pair`, while three-class supervision exposes a weak context-sensitive signal.

port has no separate evidence channel and is mostly scored as absence of direct-cause evidence.

Three-class supervision changes the geometry, but only partially. Gold `emo-context` receives the highest mean `emo-context` probability, and candidate removal produces the largest `emo-context`-probability drop for gold `emo-context` pairs. This shows that the role is learnable when exposed by the label space. Yet the signal remains fragile, as the three-class RWC-Fusion model predicts `emo-context` for only 21.3% of gold `emo-context` pairs, while 69.1% are predicted as `non-pair`. Thus, `emo-context` is not simply noise or impossible to learn; exposing a context channel does not ensure that models preserve it reliably. Appendix D.1 repeats these diagnostics across text-grounded and fusion controls.

5 When Binary Pair Scores Favor Shortcut-Compatible Attributions

5.1 Natural Shortcut vs. Evidence Stress Test

We instantiate P5 with a natural stress test that does not edit utterances. The shortcut-compatible control consists of final-labeled `non-pair` candidates that are local to the target and lexically overlapping with it. The evidence-supported cases are long distance `emo-cause` pairs and original-negative `emo-context` pairs. This contrast reverses the usual alignment between evidence and easy structural cues. The unsupported control has locality

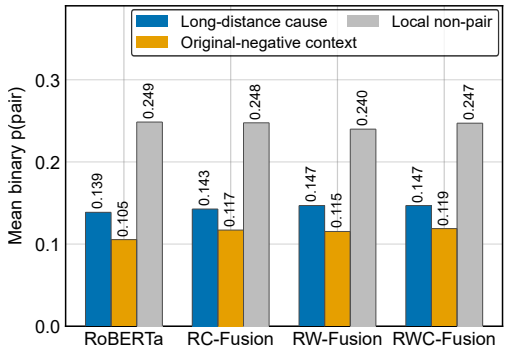

Figure 6: Binary pair scores favor shortcut-compatible `non-pair` controls over evidence-supported but structurally harder `emo-cause` and `emo-context` pairs.

and lexical overlap, while the evidence-supported cases are structurally harder.

Figure 6 shows the same ordering across binary-source controls. Local lexical `non-pair` candidates receive the highest mean pair scores, above both long-distance `emo-cause` pairs and original-negative `emo-context` pairs. The latter two subsets provide stronger explanatory evidence, but lack the shortcut-compatible locality and lexical overlap of the control. This does not show that the models ignore utterance evidence. It shows that binary pair scores can reward convenient attribution over evidence grounded emotion explanation.

This is where the proxy stops supporting emotion explanation. A high pair score can reflect convenient attribution rather than stronger evidence for why the emotion occurs. See Appendix D.3 for full results.

5.2 Matched Context Non-Pair Controls

The shortcut vs. evidence stress test shows that unsupported local candidates can score highly when they align with locality and lexical overlap. A stricter test asks whether contextual support remains separable after these cues are controlled. For each gold `emo-context` pair, we select a final-labeled `non-pair` control matched on distance, source-binary metadata, speaker relation, target emotion, candidate length, lexical overlap, and dialogue type. We then compare matched-pair score gaps under binary-source and three-class controls. This design tests whether models preserve a context signal once obvious structural and lexical explanations no longer distinguish the roles.

Figure 7 shows a consistent but narrow separation. Gold `emo-context` pairs receive slightly higher scores than their matched `non-pair` controls, with binary $p(\text{pair})$ gaps around 0.020

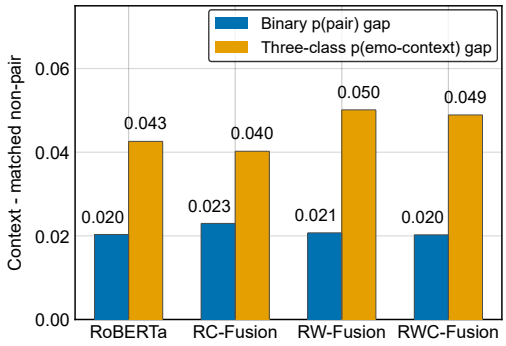

Figure 7: Matched `emo-context-non-pair` score gaps. Positive gaps indicate higher scores for gold `emo-context` pairs than matched `non-pair` controls.

- 0.023 and three-class $p(\texttt{emo-context})$ gaps around 0.040 - 0.050. The positive gaps show that `emo-context` is not simply noise or an unlearnable residual class. Their small magnitude is the diagnostic point. Once shortcut-compatible cues are matched away, the evaluated models retain only a weak signal for non-triggering discourse evidence. The validity risk is not only that unsupported local candidates can be over-ranked, but also that the boundary between contextual support and unsupported plausibility remains weakly represented.

5.3 Observable Shortcut Probes

The stress test provides one controlled contrast. We next ask whether observable structure predicts binary-source behavior more broadly. For each trained model, we fit lightweight cross-validated probes using distance, turn position, speaker relation, target and candidate emotions, dialogue metadata, length buckets, and lexical overlap. These probes exclude utterance text and multimodal content, measuring how much model behavior is recoverable from candidate-pair structure alone.

Table 4 shows that these observable features predict binary-source hard pair decisions with AUC between 0.833 and 0.855, and recover pair-score rankings with Spearman correlations between 0.462 and 0.605. This does not imply that ECPE models ignore evidence. It shows that a substantial part of binary-source behavior is reachable from structural and lexical cues alone. The validity risk is therefore not limited to one stress subset. Binary ECPE evaluation can reward structurally plausible pair decisions without requiring a stable channel for non-triggering discourse evidence.

Distance is a risk factor, not the role definition. Distance explains why locality matters, but it does not define `emo-context`. Appendix C.2 shows that

Table 4: Observable shortcut probes for binary-source behavior. AUC/Acc. evaluate hard pair decisions, and R^2/ρ evaluate pair-score prediction.

Model	AUC	Acc.	R^2	ρ
RoBERTa	$0.855_{\pm 0.007}$	$79.6_{\pm 1.1}$	$0.282_{\pm 0.058}$	$0.548_{\pm 0.063}$
RW-Fusion	$0.851_{\pm 0.025}$	$80.1_{\pm 2.7}$	$0.295_{\pm 0.058}$	$0.605_{\pm 0.054}$
RC-Fusion	$0.850_{\pm 0.020}$	$79.1_{\pm 2.0}$	$0.269_{\pm 0.030}$	$0.462_{\pm 0.114}$
RWC-Fusion	$0.833_{\pm 0.028}$	$78.8_{\pm 2.3}$	$0.263_{\pm 0.079}$	$0.562_{\pm 0.085}$

`emo-cause` concentrates at short distances, while `emo-context` shifts longer. Final file diagnostics show that 68.4% of `emo-context` pairs occur at distance ≥ 3, but 1,375 occur at shorter distances, and same distance candidates can differ in role. We therefore treat distance as a shortcut risk factor rather than an annotation rule.

Together with the stress test, this shows why binary pair scores should not be read as explanatory evidence by default. They partly track structural plausibility, while the channel for non-triggering discourse evidence remains weak. See Appendix D.4 for full proxy-only results, including source-binary and three-class targets.

6 Discussion

Scope of the claim. The results do not invalidate binary ECPE. When the task is direct-cause extraction, a pair/non-pair output remains meaningful if described as such. The problem begins when the same score is used as evidence that a model explains why the target emotion occurred. In that setting, a system may recover many direct causes while collapsing `emo-context` into `non-pair` and ranking shortcut-compatible unsupported candidates above harder evidence-supported pairs. The lesson is not that binary ECPE is broken; it is that binary pair scores have been asked to certify more than they measure.

Applied relevance. This distinction matters because deployed dialogue systems may use pair scores to produce user-facing emotion explanations. In customer-service analysis, a returned package may be the direct trigger, while prior complaints or unresolved service history explain why the frustration is severe. In healthcare, HR, education, and legal dialogue analysis, earlier disclosures, authority relations, prior feedback, or interaction history may shape emotion interpretation without being the immediate cause. In such settings, binary pair scores should be treated as direct-cause signals, not as evidence grounded emotion explanations.

How IEMO-MECP should be used. IEMO-MECP should be used as a diagnostic instrument for a label-space bottleneck. It is not a universal ontology of emotional explanation, and it does not assume that every emotion has an observable dialogue cause. The `emo-context` criterion is intentionally strict, such that it requires observable support whose removal weakens interpretation of the target emotion. This keeps `emo-context` from becoming a catch-all label for preceding utterances.

Evaluation recommendation. Future ECPE evaluations should report direct-cause extraction and emotion explanation diagnostics separately. Pair F1 is appropriate for the former, but it should not be treated as sufficient evidence for emotion explanation in dialogue. Explanation oriented systems should add role-wise metrics, boundary uncertainty analysis, shortcut vs. evidence stress tests, and error analyses that separate unsupported local plausibility from discourse evidence.

7 Related Work

Emotion-cause extraction. ECPE was introduced as a pair extraction task over emotions and causes (Xia and Ding, 2019). Subsequent ECPE work improved pair modeling through 2D representations and inter-clause interaction (Ding et al., 2020a; Wei et al., 2020), as well as alternative training or decoding formulations such as sliding-window learning, sequence labeling, and boundary-aware constraints (Ding et al., 2020b; Yuan et al., 2020; Feng et al., 2023). More recent work has explored prompting and knowledge use, fine-grained clue learning, conversational event modeling, and unified emotion-cause analysis across related tasks (Gu et al., 2024; Yu et al., 2024; Wang et al., 2024a; Yu et al., 2025). These approaches improve extraction within ECPE-style pair/non-pair label spaces. Our work instead asks which emotion explanation claims such label spaces can support.

Conversational and causal emotion analysis. Conversational emotion-cause analysis extends ECPE from documents to dialogue, speaker interaction, and conversational evidence (Poria et al., 2021; Li et al., 2023; Jeong and Bak, 2023). Causal emotion entailment further highlights the role of commonsense and discourse evidence in emotion reasoning (Zhao et al., 2023). We share this emphasis on dialogue evidence, but study a different question: *whether a binary pair score can distinguish direct causality from contextual support*.

Multimodal emotion-cause and emotion recognition. Multimodal ECPE and multimodal emotion-cause analysis incorporate text, speech, and visual evidence (Wang et al., 2023, 2024b; Li et al., 2025; Liang et al., 2025). Broader emotion-recognition work in conversations approaches context and modality from related but different angles, including multimodal fusion and semantic refinement (Shi and Huang, 2023; Zhang and Li, 2023), temporal and cross-modal interaction modeling (Nguyen et al., 2023; Yun et al., 2024), and recent work on graph-based structure, explicit evidence and clue modeling, causal prompting, and cross-space interaction (Ai et al., 2025; Zhang and Tan, 2025; Shen et al., 2025; Jing et al., 2026; Lyu et al., 2026). These works motivate the importance of context and modality. Our findings show that stronger contextual and multimodal modeling alone does not resolve a pair-label validity problem.

Context, bias, and diagnostic evaluation. Emotion-cause analysis has long considered context awareness and position bias (Li et al., 2018; Yan et al., 2021). More broadly, emotion modeling work shows that emotion-label granularity and annotation reliability affect what benchmark scores mean (Wen et al., 2023; Troiano et al., 2023). IEMO-MECP follows this diagnostic tradition, but focuses on an ECPE-specific bottleneck: *whether the inherited binary label space compresses contextual evidence needed for emotion explanation*.

8 Conclusion

Binary ECPE is useful for direct-cause extraction, but its pair scores should not be treated as sufficient evidence for evidence grounded emotion explanation. Using IEMO-MECP as a diagnostic instrument, we show that `emo-context` crosses both sides of the inherited binary boundary, concentrates near binary uncertainty, and is often collapsed into `non-pair`. A shortcut vs. evidence stress test further shows that unsupported local lexical `non-pair` controls can outrank long-distance `emo-cause` and original-negative `emo-context` pairs with stronger explanatory evidence. The lesson is not that binary ECPE is wrong; it is that binary pair scores answer a narrower question than emotion explanation requires. Future evaluations should therefore separate direct-cause extraction from evidence grounded emotion explanation in dialogue.

Limitations

IEMO-MECP is designed as a diagnostic audit of an inherited ECPE label space. Its purpose is to test whether the pair/non-pair boundary compresses a separable discourse-evidence role, rather than to serve as a universal ontology of emotion explanation. The construction and validation combine LLM-assisted first-pass labeling, dialogue-level review, targeted human refinement, additional human review of high-risk boundary transitions, a blind human reannotation audit over boundary-focused packets, and final file consistency checks. These checks provide targeted evidence for the boundary-compression analysis.

The current audit is based on IEMOCAP-style dyadic dialogues, so future work can extend the same diagnostic protocol to spontaneous, multi-party, domain-specific, and culturally diverse interactions. The blind human reannotation audit supports reproducibility in difficult boundary regions, while larger-scale independent annotation may further clarify cases where background information can be interpreted as part of a causal chain. Finally, future systems may benefit from richer pragmatic, discourse, and interactional cues to better preserve contextual support.

References

Wei Ai, Fuchen Zhang, Yuntao Shou, Tao Meng, Haowen Chen, and Keqin Li. 2025. Revisiting multimodal emotion recognition in conversation from the perspective of graph spectrum. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11):11418–11426.

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

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170. Association for Computational Linguistics.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583. Association for Computational Linguistics.

Huawen Feng, Junlong Liu, Junhao Zheng, Haibin Chen, Xichen Shang, and Qianli Ma. 2023. Joint constrained learning with boundary-adjusting for emotion-cause pair extraction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1131, Toronto, Canada. 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.

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.

Ran Jing, Geng Tu, Yice Zhang, and Ruifeng Xu. 2026. Causal-ERC: A multimodal framework with causal prompting for emotion recognition in conversations with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):31383–31391.

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):1–19.

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.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4752–4757, Brussels, Belgium. Association for Computational Linguistics.

Qiao Liang, Ying Shen, Tiantian Chen, and Lin Zhang. 2025. M3hg: Multimodal, multi-scale, and multi-type node heterogeneous graph for emotion cause triplet extraction in conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11416–11431, Vienna, Austria. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1907.11692*.

Xiaosen Lyu, Jiayu Xiong, Yuren Chen, Wanlong Wang, Xiaoqing Dai, and Jing Wang. 2026. Cross-space synergy: A unified framework for multimodal emotion recognition in conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(29):24226–24234.

Cam-Van Thi Nguyen, Anh-Tuan Mai, The-Son Le, Hai-Dang Kieu, and Duc-Trong Le. 2023. Conversation understanding using relational temporal graph neural networks with auxiliary cross-modality interaction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15154–15167, Singapore.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark,

Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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.

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.

Enrica Troiano, Laura Oberländer, and Roman Klinger. 2023. Dimensional modeling of emotions in text with appraisal theories: Corpus creation, annotation reliability, and prediction. *Computational Linguistics*, 49(1):1–72.

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.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023. 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.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181. Association for Computational Linguistics.

Jintao Wen, Geng Tu, Rui Li, Dazhi Jiang, and Wenhua Zhu. 2023. Learning more from mixed emotions: A label refinement method for emotion recognition in conversations. *Transactions of the Association for Computational Linguistics*, 11:1485–1499.

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.

Hanqi Yan, Lin Gui, Gabriele Pergola, and Yulan He. 2021. Position bias mitigation: A knowledge-aware graph model for emotion cause extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3364–3375. Association for Computational Linguistics.

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.

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.

Chaofa Yuan, Chuang Fan, Jianzhu Bao, and Ruifeng Xu. 2020. Emotion-cause pair extraction as sequence labeling based on a novel tagging scheme. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3568–3573. 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.

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.

Xiaoheng Zhang and Yang Li. 2023. A cross-modality context fusion and semantic refinement network for emotion recognition in conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13099–13110, Toronto, Canada. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):14020–14028.

A Diagnostic Instrument and Annotation Criteria

This section documents the construction controls for IEMO-MECP while keeping the focus on the evaluation-validity claim of the paper. The audit is not intended to replace existing ECPE benchmarks. It fixes the inherited candidate space, preserves source-binary metadata, and adds pair-role labels so that boundary compression can be inspected directly.

A.1 Diagnostic Workflow

Figure 8 summarizes the construction-side logic. The workflow starts from the label-space bottleneck where binary ECPE scores place direct causes, contextual support, and unsupported candidates on one pair/non-pair axis. The boundary-diagnosis stage then targets regions where that axis is likely to hide non-triggering discourse evidence. Pair-role auditing separates the same candidate pairs into `emo-cause`, `emo-context`, and `non-pair` while retaining the original source-binary label for each pair. The resulting data allow binary and role-aware views to be compared over the same lower-triangular dialogue-pair space.

The workflow is diagnostic rather than promotional. Its purpose is to make boundary compression measurable, such that if `emo-context` crosses the inherited boundary, concentrates near binary uncertainty, or collapses back into `non-pair` under model scoring, these patterns can be interpreted as evidence about what binary pair scores do and do not measure.

A.2 Label Provenance and Targeted Human Judging

The final labels are not unchecked LLM outputs. The LLM-assisted stage supplies structured role proposals and rationales, but the released labels retain explicit provenance for how each candidate enters the final files. Table 5 gives the final-file provenance distribution. The key diagnostic point is that all final `emo-context` pairs have non-default provenance, and all are associated with targeted human judging. Thus, `emo-context` is not introduced by the default lower-triangular `non-pair` fill, by a distance rule, or by an unreviewed automatic pass.

Human judging is targeted rather than uniform reannotation of every candidate pair. This design matches the validity question. Review is concentrated where the main claims depend on boundary distinctions. Original positives proposed away from direct causality, original negatives proposed as explanatory, targets whose inherited cause coverage becomes inconsistent after refinement, and candidate regions where `emo-context` could otherwise be confused with proximity or weak causality. The "reviewed pairs" column in Table 6 counts rows in the review files. The final-provenance columns count released labels whose final source points to that review region. These counts need not match exactly because later consistency decisions and overlapping review regions can supersede earlier rows.

A.3 Blind Human Re-annotation Audit

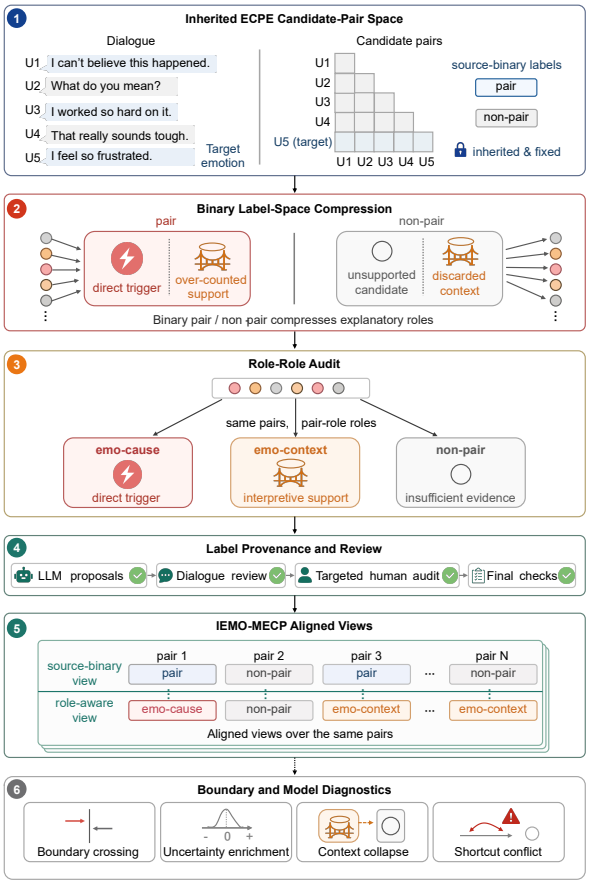

Figure 8: Construction-side diagnostic workflow for IEMO-MECP. The candidate space stays fixed, while role labels and source-binary metadata make boundary compression measurable.

Audit design. The targeted human-judge inventory in Appendix A.2 documents where final labels receive explicit human provenance. We additionally test whether the central pair-role boundary can be recovered under independent blind judgment. The audit uses target-centered packets where each packet contains one target emotion utterance and

Table 5: Final pair-role label provenance. Counts are final released labels.

Provenance source	Total	emo-cause	emo-context	non-pair
Default lower-triangular fill	70,649	0	0	70,649
LLM-assisted retained labels	3,631	3,631	0	0
Targeted human judge	11,427	5,004	4,346	2,077
Human refinement actions	368	0	0	368
Final total	86,075	8,635	4,346	73,094

Table 6: Targeted human-judge inventory. Reviewed regions are boundary cases where source labels, LLM proposals, and consistency checks are likely to diverge.

Human-judge region	Reviewed	Manual decision in review file			Final labels with this provenance		
	pairs	emo-cause	emo-context	non-pair	emo-cause	emo-context	non-pair
Multi-cause consistency review	2,678	2,428	250	0	2,428	201	0
Single-cause consistency review	2,082	2,082	0	0	2,001	0	0
Nearby-context consistency review	448	0	448	0	0	448	0
Original negative proposed as emo-cause	2,517	327	118	2,072	213	118	2,047
Original negative proposed as emo-context	2,975	0	2,969	6	0	2,969	6
Original positive proposed as emo-context	724	292	432	0	78	432	0
Original positive proposed as non-pair	2,138	2,016	79	43	96	60	24
Source-positive recovery after lost cause coverage	306	188	118	0	188	118	0

Table 7: Boundary-focused packets used in the blind human reannotation audit. Packet settings and role contrasts are used only for sampling and are not shown to annotators.

Packet setting	Role contrast	Packets	Candidate composition	Judgments
Two-role	emo-cause / emo-context	40	40 emo-cause + 40 emo-context	80
Two-role	emo-context / non-pair	40	40 emo-context + 40 non-pair	80
Two-role	emo-cause / non-pair	40	40 emo-cause + 40 non-pair	80
Three-role	emo-cause / emo-context / non-pair	30	30 emo-cause + 30 emo-context + 30 non-pair	90
Total	All contrasts	150	110 emo-cause + 110 emo-context + 110 non-pair	330

two or three candidate utterances. Annotation remains pair-level, with one candidate-to-target judgment per row.

The audit is boundary-focused rather than distribution-matching. We sample packets without reusing the same target utterance, prioritize high-overlap or distance-matched boundary cases, and include candidates from targeted human-judge regions when available. The hidden final labels are balanced across emo-cause, emo-context, and non-pair, but these labels are used only for audit construction. Table 7 summarizes the packet design.

Blind annotation setting. Three annotators independently label all candidate-to-target judgments. They see the dialogue context, the target emotion utterance, the target emotion label, and the candidate utterance. They do not see the final pair-role label, the source-binary label, the label provenance, the LLM proposal or rationale, or the packet type. Agreement therefore measures independent recovery of the role boundary rather than agreement with a visible proposal.

Agreement pattern. Table 8 shows that the role boundary remains reproducible even on packets selected to stress difficult distinctions. The lowest agreement appears for emo-context, as expected for the boundary role, but it still receives stable majority behavior under blind annotation. The result supports the use of emo-context as a recoverable discourse-evidence role rather than as an unchecked artifact of the construction pipeline.

Diagnostic interpretation. Table 9 shows that the disagreement structure is consistent with the role hierarchy in Section 2. Most ambiguity is between direct-cause evidence and non-triggering discourse evidence, not between emo-context and ordinary non-pair evidence. The audit therefore does not claim that every boundary case is trivial. It shows that the difficult boundary is structured, where annotators mostly disagree over whether discourse evidence is directly causal or contextual, rather than over whether emo-context is merely unsupported proximity.

Table 8: Agreement in the blind human reannotation audit. Raw agreement is averaged over annotator pairs. Class-conditioned columns use the hidden final label only for analysis. These labels are not visible during annotation.

Metric	Overall	emo-cause	emo-context	non-pair
Raw pairwise agreement	0.821	0.811	0.767	0.893
Mean pairwise Cohen's κ	0.732	–	–	–
Pairwise Cohen's κ range	0.696–0.758	–	–	–
Krippendorff's α	0.732	–	–	–
3/3 unanimous judgments	245 / 330	83 / 110	75 / 110	87 / 110
2/3 majority judgments	78 / 330	33 / 110	32 / 110	13 / 110

Table 9: Disagreement patterns in the blind human reannotation audit. Counts are computed over non-unanimous judgments.

Disagreement type	Count	%
emo-cause vs. emo-context	52	61.2
emo-context vs. non-pair	19	22.4
emo-cause vs. non-pair	7	8.2
All three labels different	7	8.2
Total	85	100.0

A.4 LLM-Assisted Draft-to-Final Changes

Table 10 compares the OpenAI Codex draft layer, produced with the `gpt-5.5 xhigh` model/reasoning setting, with the final labels after targeted human audit. The draft is defined over the full candidate space, where pairs not elevated by the LLM-assisted stages are counted as draft `non-pair`. The "Changed" column therefore measures draft-to-final movement after audit, not an LLM error rate.

The movement pattern is diagnostic. Original positives change more often because the audit restores direct-cause coverage and separates over-broad positives from non-triggering support. Original negatives remain comparatively stable, while a targeted subset is elevated to `emo-context` or `emo-cause` when the dialogue supplies evidence. This supports the intended use of IEMO-MECP, where the audit refines high-risk boundary regions while preserving the inherited binary task structure.

A.5 Final-File Diagnostic Checks

Table 11 provides the main file-level sanity checks behind the diagnostic audit. The first row addresses task preservation, such that most source-positive pairs remain `emo-cause`, and most source-negative pairs remain `non-pair`. This supports the claim that IEMO-MECP refines binary ECPE rather than replacing it with an unrelated labeling scheme.

The second row is the central boundary-compression evidence. `emo-context` appears among both source positives and source negatives. This two-sided pattern matters because it argues against two simpler interpretations, where `emo-context` is neither a set of missing positives, nor a weaker version of `emo-cause`. Instead, the same discourse role is forced onto both sides of the inherited binary boundary.

The remaining rows guard against common failure modes in interpreting the labels. Explicit provenance for all `emo-context` pairs shows that the class was not produced by default negative filling. The self-pair row shows that same-turn candidates are not automatically treated as direct causes; self-pairs require observable causal or appraisal evidence. The distance row shows that `emo-context` is distance-skewed but not distance-defined. Together, these checks support the main-paper claim that `emo-context` is a discourse-evidence role rather than a construction artifact.

A.6 Boundary Examples

Table 12 illustrates the same discourse role moving through both sides of the inherited boundary. Positive-to-context cases show over-causalized support. The candidate helps interpret the target emotion, but the target utterance already contains the main emotional expression or appraisal. At the same time, negative-to-context cases show discarded discourse evidence where the inherited source label is negative, yet the candidate still makes the target emotion easier to interpret.

These examples make the linguistic role of `emo-context` explicit. The positive-to-context cases should not be counted as direct triggers merely because they are helpful. The negative-to-context cases should not be discarded merely because they are not triggers. This is the same validity problem studied in the main text, where ***binary ECPE can keep many direct causes while still compressing evidence needed for grounded explanation in emotion***.

Table 10: OpenAI Codex draft labels compared with final labels after targeted human audit. Draft `non-pair` includes pairs not elevated by the LLM-assisted stages.

Scope	Pairs	OpenAI Codex draft			Final after human audit			Changed
		`emo-cause`	`emo-context`	`non-pair`	`emo-cause`	`emo-context`	`non-pair`	
Overall	86,075	8,579	3,699	73,797	8,635	4,346	73,094	5,313 (6.17%)
Original positives	9,145	6,062	724	2,359	8,316	805	24	2,688 (29.39%)
Original negatives	76,930	2,517	2,975	71,438	319	3,541	73,070	2,625 (3.41%)

Table 11: Extended final-file diagnostics supporting the boundary-compression analysis.

Diagnostic	Final-file evidence	Interpretation
Binary task is preserved	8,316/9,145 original positives remain `emo-cause` (90.9%); 73,070/76,930 original negatives remain `non-pair` (95.0%).	The role-aware view refines the source ECPE boundary rather than replacing it with an unrelated task.
Context crosses the inherited boundary	805/9,145 original positives (8.8%) and 3,541/76,930 original negatives (4.6%) become `emo-context`.	`emo-context` is not only a missing-positive phenomenon and not only a weakened cause label.
Context labels have explicit provenance	4,346/4,346 final `emo-context` pairs have explicit non-default provenance in the released label files.	The context class is not produced by the default `non-pair` fill or by a distance-only rule.
Self-pairs are not automatic causes	Self-pairs split into 2,794 `emo-cause` and 4,639 `non-pair` cases, with 0 `emo-context` cases.	A self-pair is accepted as `emo-cause` only when the target utterance itself contains sufficient causal or appraisal evidence.
Distance is a risk factor, not a rule	2,971/4,346 `emo-context` pairs occur at distance ≥ 3 (68.4%), but 1,375 occur at shorter distances.	`emo-context` is distance-skewed, yet still requires semantic evidence specific to the target emotion.

B Formalizing Binary Boundary Compression

The main text treats binary ECPE as a projection from a role-aware evidential space to a two-label decision space. Here we state the implication explicitly. Let $\mathcal{R} = \{\texttt{emo-cause}, \texttt{emo-context}, \texttt{non-pair}\}$, where `emo-cause` is direct-cause evidence, `emo-context` is non-triggering explanatory evidence, and `non-pair` is insufficient evidence. Binary ECPE observes only $y^{\text{bin}} \in \{0, 1\}$, so any binary evaluation induces a projection $\pi : \mathcal{R} \to \{0, 1\}$.

For direct-cause extraction, $\pi(\texttt{emo-cause}) = 1$ and $\pi(\texttt{non-pair}) = 0$. The remaining role, `emo-context`, is non-causal but explanatorily relevant. If $\pi(\texttt{emo-context}) = 1$, the projection preserves explanatory relevance but loses causal status by evaluating contextual support as a cause. If $\pi(\texttt{emo-context}) = 0$, the projection preserves non-causality but loses explanatory relevance by evaluating contextual support as irrelevant. Therefore, no binary projection that preserves the usual direct-cause interpretation can also preserve the distinction among direct causality, non-triggering explanatory support, and insufficient evidence.

This does not imply that binary ECPE is invalid. It implies that binary ECPE is valid for a narrower measurement target, which is direct-cause extraction. When the same binary score is interpreted as evidence grounded emotion explanation, the projection becomes under-specified. The empirical diagnostics in the main paper test whether this formal compression appears in the inherited ECPE candidate space and in model behavior.

C Training and Implementation Details

This section fixes the implementation setting for the role-level diagnostics. The goal is not leaderboard optimization. It is to make clear which splits, model recipes, environments, and construction-support tools underlie the reported comparisons.

C.1 Final IEMO-MECP Split Statistics

Table 13 reports the released split structure and final pair-role counts. Raw dialogues are unique `raw_doc_id` units; chunks are the chunked dialogue instances in the split files; utterances are counted within those chunks. The table also confirms that the benchmark keeps the original lower-triangular candidate-pair space while evaluating three-class supervision over `emo-cause`, `emo-context`, and `non-pair`.

The split statistics explain the evaluation choices

Table 12: Representative transitions from the original binary space to the final pair role-aware view.

Transition	Target utterance	Candidate utterance	Role rationale
Positive → context	The campus is beautiful. I'm so excited.	It's safe around there.	The candidate adds situational background, but the target already expresses the emotion.
Positive → context	I feel really bad for his family, you know.	Yeah, he kind of did a lot of things, but we acted together, sometimes.	The candidate adds background about shared responsibility, but the target already expresses the emotion.
Negative → context	This is ridiculous. I seriously do not understand why you think these automated systems are supposed to work for anybody.	All right. I'm sorry, sir that you've been having a hard time.	The candidate supports interpretation of the target frustration, but it does not itself trigger the emotion.
Negative → context	I didn't know him, Tricia, but uh- I heard that he was a great man.	Turned around for the worst, you know.	The candidate supplies interpretive background for the target's sadness, but it does not directly cause the emotion.

Table 13: Released split statistics for IEMO-MECP.

Statistic	Train	Valid	Test	Overall
Pairs	53,323	14,488	18,264	86,075
`emo-cause`	5,499	1,342	1,794	8,635
`emo-context`	2,630	719	997	4,346
`non-pair`	45,194	12,427	15,473	73,094
Avg. chunks/dialog	2.71	2.83	3.06	2.80
Avg. utterances/chunk	17.64	17.99	17.08	17.57
Avg. pairs/chunk	205.09	213.06	192.25	203.49

Table 14: Architecture-specific training recipes.

Model	Optimizer	Distinct settings
MECPE-2step	Adam	step 1: batch 8, lr 5×10^{-3}, 15 iters; step 2: batch 64, lr 5×10^{-3}, 12 iters
HiLo	AdamW	BERT lr 1×10^{-5}, task lr 1×10^{-4}, batch 4, 20 epochs
M^3HG	AdamW	train batch 8 with gradient accumulation 2; eval batch 8; lr 5×10^{-5}; 50 epochs

in the main paper. The overall class distribution is strongly skewed toward `non-pair`, which motivates macro F1 and class-wise F1 in the main evaluation. At the same time, `emo-context` contains 4,346 candidate pairs overall, so it is not merely an anecdotal boundary category. The same three-role structure is present in train, validation, and test splits, allowing `emo-context` behavior to be evaluated under the same candidate-space assumptions as `emo-cause` and `non-pair`.

C.2 IEMO-MECP Distance Distributions

Figure 9 shows that `emo-context` is shifted toward longer distances than `emo-cause` across splits. This helps explain why contextual support can collapse toward `non-pair` in models biased toward local causes.

However, distance is not the role definition. The role distributions overlap, and shorter-distance `emo-context` pairs remain. Distance is therefore a shortcut risk factor, not an annotation rule. The semantic criterion is unchanged, such that `emo-context` must provide non-triggering support whose removal weakens interpretation of the target emotion.

C.3 Optimization Settings

All runs optimize the three-class pair-role task with class-weighted cross-entropy. Class weights are computed from the training split for the active task. RoBERTa, WavLM, CLIP, RW-Fusion, RC-Fusion, and RWC-Fusion share a single baseline recipe as follows. AdamW with learning rate 2×10^{-5}, training batch size 16, evaluation batch size 32, 10 epochs, warmup ratio 0.1, and max length 160. Only the enabled modalities differ. Checkpoints are selected by validation pair-role macro F1.

Table 14 isolates settings that vary for the stronger model families. The shared baseline recipe fixes the simple unimodal and fusion controls, while the table reports only architecture-specific choices for MECPE-2step, HiLo, and M^3HG. This separation helps interpret model-family differences without turning the appendix into a hyperparameter search.

C.4 Model Size and Parameter Counts

Table 15 reports the parameter counts for the evaluated model families. For settings that use precomputed WavLM or CLIP features, we report the trainable predictor parameters and exclude fixed feature tables. MECPE-2step counts also exclude fixed

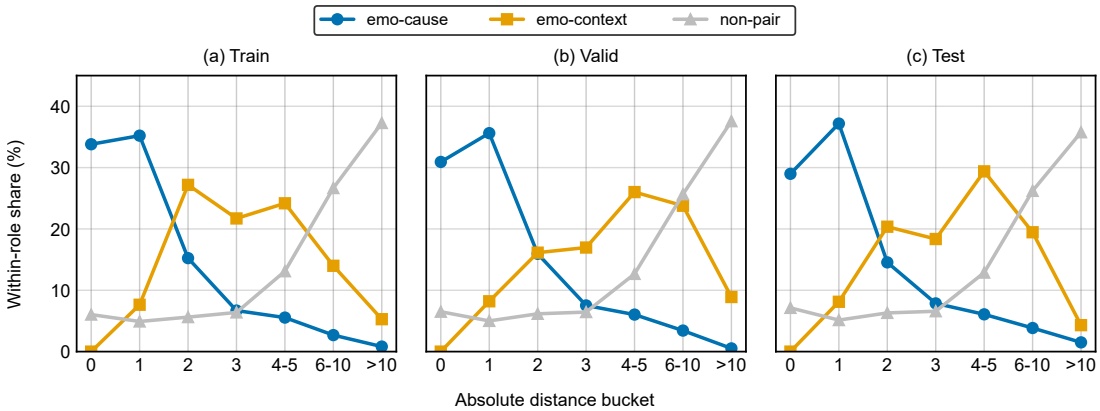

Figure 9: Role-conditioned absolute-distance distributions by split. `emo-context` shifts toward longer distances than `emo-cause`, but distance buckets overlap across roles.

Table 15: Model size summary. Counts are reported in millions of parameters. For feature-based audio/video settings and MECPE-2step, fixed feature or embedding tables are excluded as noted in the text.

Model setting	**Parameters**
RoBERTa (T)	124.65M
WavLM feature MLP (A)	0.39M trainable
CLIP feature MLP (V)	0.39M trainable
RW / RC / RWC-Fusion	125.83M / 125.83M / 127.02M
MECPE-2step, T to T+A+V	1.82M–2.77M trainable
HiLo (T+A+V)	153.28M
M^3HG, T to T+A+V	171.61M–181.85M

embedding or feature tables. Fusion and graph-model rows report the instantiated model sizes for the corresponding modality settings.

C.5 Compute Environment

The reported experiments were run on the server summarized in Table 16. The core multimodal pipeline uses the Python/PyTorch stack listed there, while MECPE-2step is executed through a separate TensorFlow-compatible launcher.

The experimental budget consists of 54 seeded three-class pair-role runs, covering 18 model-modality settings over three seeds, and 18 seeded binary-source diagnostic runs. This yields 72 one-GPU training/evaluation runs. Each run used one A100 80GB GPU. Summing the per-job durations recoverable from launcher logs gives approximately 65.3 A100 GPU-hours, equivalent to about 32.6 fully utilized hours on the two-A100 server.

C.6 AI-Assistance Disclosure

OpenAI Codex with the model and reasoning setting reported as `gpt-5.5 xhigh` was used for LLM-assisted construction support. The outputs were used as proposal, review, and verification signals under the safeguards described in Section 3; they were not treated as unchecked gold annotations or unverified scientific claims. The recorded Codex usage for the annotation and dataset-construction workflow totaled approximately 509 million tokens. This disclosure is included because IEMO-MECP is positioned as a diagnostic audit rather than as a fully human-reannotated corpus from scratch.

Table 16: Compute environment and key package versions.

Item	Value
Operating system	Linux 5.15.0-1045-nvidia
GPUs	2 × NVIDIA A100-SXM4-80GB
Driver	535.216.03
Python	3.12.12
PyTorch	2.10.0+cu128
transformers	5.7.0
numpy	1.26.4
pandas	3.0.2
matplotlib	3.10.9
scikit-learn	1.8.0
tqdm	4.67.3
PyYAML	6.0.3
seaborn	0.13.2
MECPE-2step	separate TensorFlow
launcher	2.15-compatible environment

D Boundary Evidence and Cross-Family Diagnostics

This section expands the model-side diagnostics behind the main text. The central question is whether the same validity pattern persists across model families and controls. It is found that binary ECPE preserves useful direct-cause decisions, but its score geometry and error patterns remain weak for non-triggering discourse evidence.

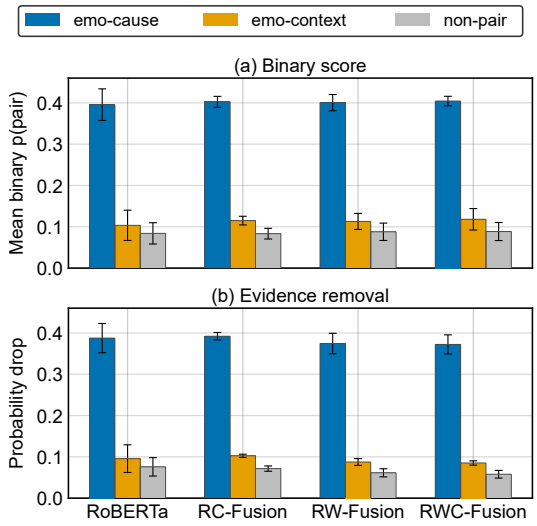

Figure 10: Binary-axis and evidence-removal diagnostics across text-grounded and fusion controls. Binary-source scores keep `emo-context` closer to `non-pair` than to `emo-cause`.

D.1 Binary-Axis and Evidence-Removal Diagnostics

Figure 10 repeats the binary-axis and candidate-removal diagnostics across text-grounded and fusion controls. The binary-axis component asks where final `emo-cause`, `emo-context`, and `non-pair` pairs fall under original binary supervision. The candidate-removal component asks how much the relevant score drops when the candidate utterance is removed.

The finding matches the main RWC-Fusion diagnostic. Gold `emo-context` receives binary pair scores closer to `non-pair` than to `emo-cause`, which indicates that binary supervision lacks a stable context channel. Candidate removal also affects `emo-cause` more strongly than `emo-context` under binary-source scoring. The model-side signal is therefore not simply that `emo-context` is difficult; it is that binary pair scores are more sensitive to direct-cause evidence than to non-triggering support.

D.2 Cross-Family Confusion Matrices

Figure 11 shows mean 3×3 pair-role confusion matrices for the evaluated model families under three-class supervision. Cells are row-normalized, so each row describes where a gold role is predicted rather than how frequent that role is in the dataset.

The confusion matrices support the role-wise conclusion in the main text. Context collapse is not confined to one baseline or one modality setting. Across model families, gold `emo-context` is frequently absorbed into `non-pair`, even when the same systems identify many `emo-cause` and `non-pair` examples more reliably. This makes the error direction part of the diagnostic evidence. Models do not merely miss a hard class; they tend to treat contextual support as absence of pair evidence.

D.3 Full Shortcut-vs-Evidence Stress Test

Table 17 expands the natural shortcut-vs-evidence stress test from Section 5. The local lexical `non-pair` subset is not adversarially edited. It consists of naturally occurring candidates that are close to the target and lexically overlapping with it, but unsupported under the final role label. The evidence-supported subsets have weaker shortcut alignment, where long-distance `emo-cause` pairs are structurally non-local, and original-negative `emo-context` pairs come from the inherited negative side.

The ordering is stable across binary-source text-grounded and fusion families. Local lexical `non-pair` controls receive higher mean pair scores and higher predicted-pair rates than both evidence-supported subsets. The distance and lexical-overlap columns explain why this is a shortcut-conflict test. The unsupported control is close and overlapping, whereas the evidence-supported subsets are farther and less lexically aligned. The table therefore supports the main claim that binary pair scores can be inflated by shortcut-compatible structure.

D.4 Full Proxy-Only Behavior Prediction

Table 18 reports the broader proxy-only probe. Observable-only features include structural, dialogue, speaker, emotion, position, length, and lexical-overlap variables, but not utterance text or multimodal content. The source-binary variant adds inherited binary metadata. The probe predicts trained-model outputs, not gold labels. High values therefore indicate that model decision or score behavior is recoverable from shortcut-compatible observable structure.

The binary-source rows show that observable-only features recover hard pair decisions with high AUC and recover pair-score rankings with moderate to strong correlations. The three-class rows show a related issue for context behavior, where context decisions and context probabilities are also partly predictable from structural variables. This does not imply that the ECPE models ignore utter-

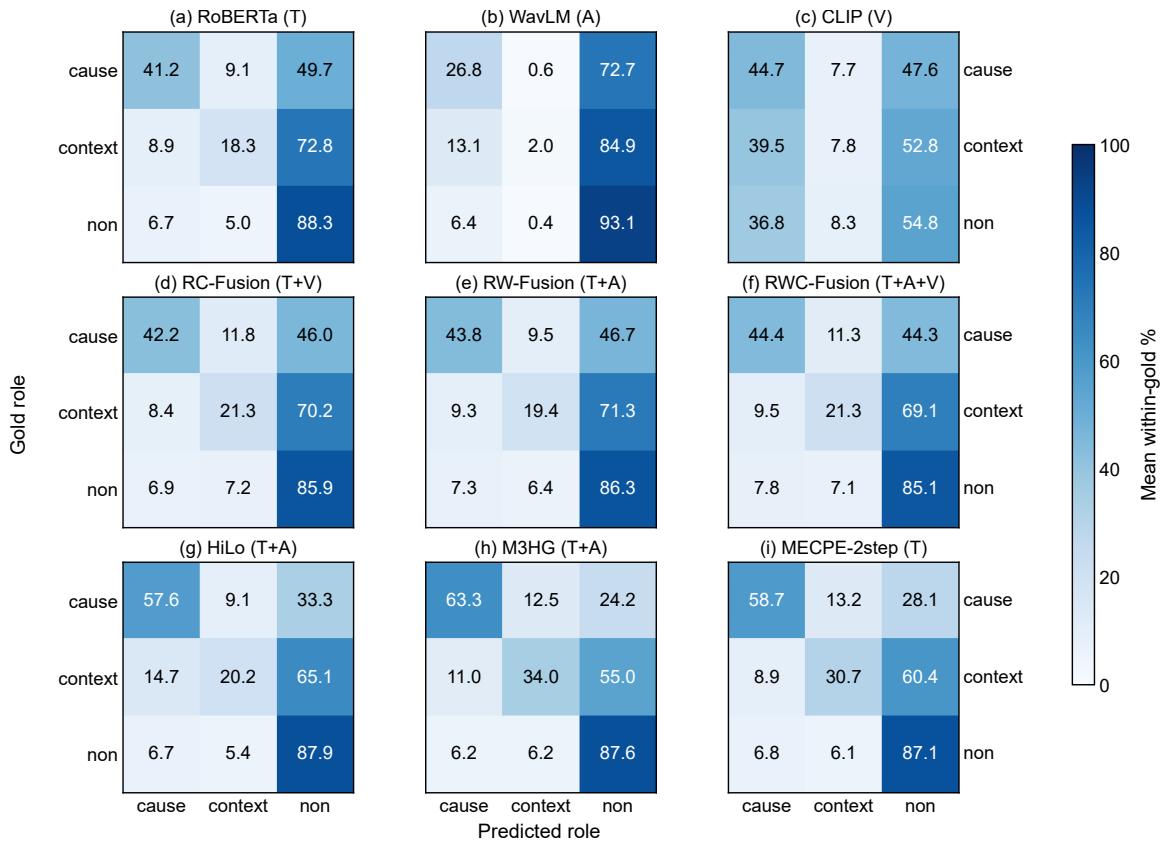

Figure 11: Row-normalized pair-role confusion matrices under three-class supervision. Across model families, the dominant `emo-context` error is collapse into `non-pair`.

Table 17: Natural shortcut-vs-evidence stress-test results for binary-source controls. Values are mean$_{\pm \text{std}}$ over seeds; n, distance, and lexical overlap are fixed by subset definition.

Model	Stress subset	n	p(pair)	Pred-pair (%)	Dist.	Lex. overlap
RoBERTa	Local lexical `non-pair`	1489	$0.302_{\pm 0.041}$	$29.0_{\pm 3.5}$	0.3	0.78
	Long-distance `emo-cause`	92	$0.139_{\pm 0.033}$	$11.6_{\pm 1.7}$	9.6	0.07
	Original-negative `emo-context`	854	$0.105_{\pm 0.037}$	$7.5_{\pm 1.1}$	4.5	0.07
RC-Fusion	Local lexical `non-pair`	1489	$0.302_{\pm 0.012}$	$28.9_{\pm 2.6}$	0.3	0.78
	Long-distance `emo-cause`	92	$0.143_{\pm 0.028}$	$11.6_{\pm 2.7}$	9.6	0.07
	Original-negative `emo-context`	854	$0.117_{\pm 0.011}$	$9.4_{\pm 2.3}$	4.5	0.07
RW-Fusion	Local lexical `non-pair`	1489	$0.292_{\pm 0.018}$	$28.3_{\pm 1.4}$	0.3	0.78
	Long-distance `emo-cause`	92	$0.147_{\pm 0.020}$	$10.5_{\pm 1.7}$	9.6	0.07
	Original-negative `emo-context`	854	$0.115_{\pm 0.019}$	$7.2_{\pm 1.6}$	4.5	0.07
RWC-Fusion	Local lexical `non-pair`	1489	$0.300_{\pm 0.023}$	$29.4_{\pm 2.5}$	0.3	0.78
	Long-distance `emo-cause`	92	$0.147_{\pm 0.037}$	$11.6_{\pm 4.1}$	9.6	0.07
	Original-negative `emo-context`	854	$0.119_{\pm 0.024}$	$8.7_{\pm 1.6}$	4.5	0.07

ance evidence. It shows that score behavior is substantially aligned with shortcut-compatible structure, which is why binary pair scores should not be exposed as faithful evidence grounded emotion explanations without role-level diagnostics.

E Full Family-Level Results

This section preserves the full model evidence behind the compact main text results.

E.1 Calibration Baselines for Role-Wise Results

Table 19 provides simple calibration baselines for the role-wise results in Table 3. These baselines are diagnostic rather than competitive. They clarify whether the three-class pair-role results can be explained by majority-class behavior, distance structure, or the inherited binary output space. The distance-bucket majority baseline collapses to always-`non-pair` behavior, and its high `non-pair`

Table 18: Proxy-only prediction of trained-model behavior. Values are mean$_{\pm \mathrm{std}}$ over seeds. AUC and accuracy are reported for hard predictions; R^2 and Spearman ρ are reported for score regression. Dashes mark metrics not defined for the target type.

Source	Target	Features	Model	AUC	Acc.	R^2	ρ
Binary	hard pair	Obs.	RoBERTa	$0.855_{\pm 0.007}$	$79.6_{\pm 1.1}$	-	-
			RC-Fusion	$0.850_{\pm 0.020}$	$79.1_{\pm 2.0}$	-	-
			RW-Fusion	$0.851_{\pm 0.025}$	$80.1_{\pm 2.7}$	-	-
			RWC-Fusion	$0.833_{\pm 0.028}$	$78.8_{\pm 2.3}$	-	-
	hard pair	Obs.+src	RoBERTa	$0.858_{\pm 0.006}$	$81.0_{\pm 1.0}$	-	-
			RC-Fusion	$0.853_{\pm 0.021}$	$80.2_{\pm 2.1}$	-	-
			RW-Fusion	$0.854_{\pm 0.025}$	$81.4_{\pm 2.7}$	-	-
			RWC-Fusion	$0.836_{\pm 0.029}$	$80.2_{\pm 2.2}$	-	-
Binary	$p(\text{pair})$	Obs.	RoBERTa	-	-	$0.282_{\pm 0.058}$	$0.548_{\pm 0.063}$
			RC-Fusion	-	-	$0.269_{\pm 0.030}$	$0.462_{\pm 0.114}$
			RW-Fusion	-	-	$0.295_{\pm 0.058}$	$0.605_{\pm 0.054}$
			RWC-Fusion	-	-	$0.263_{\pm 0.079}$	$0.562_{\pm 0.085}$
	$p(\text{pair})$	Obs.+src	RoBERTa	-	-	$0.304_{\pm 0.057}$	$0.555_{\pm 0.062}$
			RC-Fusion	-	-	$0.293_{\pm 0.031}$	$0.472_{\pm 0.116}$
			RW-Fusion	-	-	$0.321_{\pm 0.059}$	$0.615_{\pm 0.056}$
			RWC-Fusion	-	-	$0.286_{\pm 0.081}$	$0.571_{\pm 0.086}$
Three-class	hard context	Obs.	RoBERTa	$0.855_{\pm 0.061}$	$77.2_{\pm 5.6}$	-	-
			RC-Fusion	$0.830_{\pm 0.051}$	$74.7_{\pm 4.7}$	-	-
			RW-Fusion	$0.844_{\pm 0.043}$	$75.7_{\pm 3.0}$	-	-
			RWC-Fusion	$0.833_{\pm 0.057}$	$74.9_{\pm 5.0}$	-	-
	hard context	Obs.+src	RoBERTa	$0.855_{\pm 0.061}$	$77.2_{\pm 5.6}$	-	-
			RC-Fusion	$0.830_{\pm 0.051}$	$74.7_{\pm 4.6}$	-	-
			RW-Fusion	$0.843_{\pm 0.043}$	$75.7_{\pm 3.0}$	-	-
			RWC-Fusion	$0.832_{\pm 0.057}$	$74.9_{\pm 5.1}$	-	-
Three-class	$p(\text{context})$	Obs.	RoBERTa	-	-	$0.244_{\pm 0.153}$	$0.537_{\pm 0.154}$
			RC-Fusion	-	-	$0.229_{\pm 0.124}$	$0.486_{\pm 0.121}$
			RW-Fusion	-	-	$0.219_{\pm 0.135}$	$0.510_{\pm 0.134}$
			RWC-Fusion	-	-	$0.258_{\pm 0.163}$	$0.547_{\pm 0.137}$
	$p(\text{context})$	Obs.+src	RoBERTa	-	-	$0.243_{\pm 0.153}$	$0.537_{\pm 0.154}$
			RC-Fusion	-	-	$0.229_{\pm 0.125}$	$0.486_{\pm 0.122}$
			RW-Fusion	-	-	$0.218_{\pm 0.135}$	$0.510_{\pm 0.134}$
			RWC-Fusion	-	-	$0.258_{\pm 0.163}$	$0.546_{\pm 0.137}$

F1 but zero `emo-cause` and `emo-context` F1 confirms why macro and role-wise F1 are more informative than majority-class behavior.

The distance-only balanced logistic regression obtains non-zero `emo-context` F1, showing that `emo-context` contains learnable distance regularity rather than behaving as random noise. However, the binary-source predictor remapped into the pair-role space still has no `emo-context` channel, because predicted pairs can only be projected to `emo-cause` and predicted non-pairs to `non-pair`. The M^{3}HG T+A reference setting gives the strongest macro F1 among these rows while preserving a non-zero `emo-context` channel and high `non-pair` F1. These calibration results support the main text interpretation, which is that the three-class pair-role results are not explained by majority behavior or a simple distance heuristic, and the inherited binary output space cannot express non-triggering contextual support as a separate role.

E.2 Source-Binary Remapping

Table 20 maps three-class predictions back to the inherited binary ECPE task. "Drop `emo-context`" evaluates only non-context predictions under the source-binary mapping. "`emo-context` → pair" treats predicted `emo-context` as source-positive. "`emo-context` → `non-pair`" treats predicted `emo-context` as source-negative.

The remapping results support the compatibility claim in the main text. Treating predicted `emo-context` as source-positive generally hurts text-grounded settings, so the role should not be collapsed into a simple positive expansion. In contrast, dropping predicted `emo-context` or mapping it to `non-pair` preserves or improves source-binary Pair F1 for the text-grounded and fusion controls. Separating `emo-context` can therefore sharpen the

Table 19: Calibration baselines for role-wise evaluation on the test split. F1 values are percentages. The M^{3}HG T+A row is the main text reference row.

Method	Macro F1	emo-cause F1	emo-context F1	non-pair F1
Distance-bucket majority (= always non-pair)	30.58	0.00	0.00	91.73
Distance-only balanced logistic regression	48.31	47.40	21.99	75.53
Binary-source predictor → pair-role remap	43.35	39.32	0.00	90.71
M^{3}HG T+A	57.95	56.73	26.84	90.30

Table 20: Source-binary Pair F1 after remapping three-class predictions. Values are mean$_{\pm \mathrm{std}}$ over seeds 42, 123, and 456.

Model	Mod.	Original binary	Drop emo-context	emo-context → pair	emo-context → non-pair
RoBERTa	T	$37.60_{\pm 1.05}$	$40.87_{\pm 0.34}$	$36.70_{\pm 0.48}$	$39.05_{\pm 0.06}$
WavLM	A	$28.24_{\pm 0.22}$	$27.67_{\pm 0.24}$	$27.48_{\pm 0.30}$	$27.58_{\pm 0.23}$
CLIP	V	$16.43_{\pm 0.94}$	$19.06_{\pm 0.62}$	$18.71_{\pm 0.15}$	$18.71_{\pm 0.33}$
RW-Fusion	T+A	$37.86_{\pm 0.68}$	$41.95_{\pm 1.18}$	$36.06_{\pm 1.34}$	$40.12_{\pm 0.98}$
RC-Fusion	T+V	$37.07_{\pm 0.28}$	$42.20_{\pm 0.69}$	$36.25_{\pm 1.92}$	$39.88_{\pm 0.99}$
RWC-Fusion	T+A+V	$37.61_{\pm 0.22}$	$41.97_{\pm 1.33}$	$36.11_{\pm 0.21}$	$39.78_{\pm 0.86}$

Table 21: Full family-level test results over three seeds.

Family	Mod.	Macro F1	Acc.	emo-cause F1	emo-context F1	non-pair F1
RoBERTa	T	$48.77_{\pm 0.39}$	$79.83_{\pm 0.41}$	$40.28_{\pm 0.52}$	$17.18_{\pm 1.47}$	$88.84_{\pm 0.29}$
WavLM	A	$40.63_{\pm 0.33}$	$81.64_{\pm 0.15}$	$28.26_{\pm 0.23}$	$3.65_{\pm 0.73}$	$89.97_{\pm 0.10}$
CLIP	V	$30.19_{\pm 1.32}$	$51.26_{\pm 3.61}$	$18.50_{\pm 0.22}$	$5.20_{\pm 2.02}$	$66.86_{\pm 3.36}$
RW-Fusion	T+A	$48.54_{\pm 0.26}$	$78.47_{\pm 0.57}$	$41.37_{\pm 1.15}$	$16.34_{\pm 1.70}$	$87.92_{\pm 0.37}$
RC-Fusion	T+V	$48.38_{\pm 0.63}$	$78.12_{\pm 1.82}$	$40.98_{\pm 1.07}$	$16.40_{\pm 2.80}$	$87.76_{\pm 1.23}$
RWC-Fusion	T+A+V	$48.41_{\pm 0.50}$	$77.61_{\pm 0.68}$	$40.89_{\pm 0.63}$	$16.94_{\pm 1.74}$	$87.41_{\pm 0.52}$
MECPE -2step	T	$55.73_{\pm 0.34}$	$81.26_{\pm 0.25}$	$52.91_{\pm 0.44}$	$24.60_{\pm 0.82}$	$89.69_{\pm 0.14}$
	T+A	$55.60_{\pm 0.08}$	$79.67_{\pm 0.94}$	$52.98_{\pm 0.11}$	$25.14_{\pm 0.91}$	$88.67_{\pm 0.60}$
	T+A+V	$55.24_{\pm 0.38}$	$79.73_{\pm 0.66}$	$52.55_{\pm 0.49}$	$24.51_{\pm 1.11}$	$88.65_{\pm 0.42}$
	T+V	$55.46_{\pm 0.11}$	$80.15_{\pm 0.35}$	$53.02_{\pm 0.31}$	$24.41_{\pm 0.63}$	$88.95_{\pm 0.20}$
HiLo	T	$52.50_{\pm 0.21}$	$80.51_{\pm 0.80}$	$50.53_{\pm 0.95}$	$17.73_{\pm 0.76}$	$89.25_{\pm 0.53}$
	T+A	$52.98_{\pm 0.80}$	$81.25_{\pm 1.32}$	$51.29_{\pm 0.44}$	$17.95_{\pm 2.67}$	$89.72_{\pm 0.86}$
	T+A+V	$52.94_{\pm 0.61}$	$79.66_{\pm 1.55}$	$50.60_{\pm 0.69}$	$19.60_{\pm 2.06}$	$88.61_{\pm 0.99}$
	T+V	$51.85_{\pm 0.62}$	$82.10_{\pm 1.11}$	$48.59_{\pm 0.61}$	$16.69_{\pm 1.64}$	$90.26_{\pm 0.63}$
M^{3}HG	T	$57.84_{\pm 0.11}$	$80.88_{\pm 0.52}$	$56.34_{\pm 0.41}$	$27.73_{\pm 0.57}$	$89.46_{\pm 0.36}$
	T+A	$57.95_{\pm 0.07}$	$82.24_{\pm 0.91}$	$56.73_{\pm 0.18}$	$26.84_{\pm 0.33}$	$90.30_{\pm 0.52}$
	T+A+V	$57.76_{\pm 0.29}$	$82.66_{\pm 0.18}$	$55.52_{\pm 0.07}$	$27.10_{\pm 1.07}$	$90.67_{\pm 0.14}$
	T+V	$57.18_{\pm 0.32}$	$81.75_{\pm 0.14}$	$55.55_{\pm 0.73}$	$25.88_{\pm 1.45}$	$90.11_{\pm 0.11}$

direct-cause channel without redefining every contextual support case as a cause.

E.3 Full Model-Modality Results

Table 21 reports mean and standard deviation over three seeds for all 18 model-modality settings. T, A, and V denote text, audio, and video inputs. Rows grouped under a family name share the same architecture with different modality choices.

The full table preserves the role-wise pattern behind the compact main text table. The highest macro F1 among these rows is achieved by M^{3}HG T+A, while the highest emo-context F1 is achieved by M^{3}HG T. This separation reinforces why aggregate scores should not be read as sufficient evidence of explanation quality. Accuracy and non-pair F1 are high partly because the candidate space is dominated by non-pair, whereas emo-context F1 remains substantially lower than emo-cause and non-pair F1 across families. The full results therefore support the same conclusion as the main text, where current systems improve direct-cause extraction more reliably than contextual support.

F Annotation Instructions and Visibility

This section reports the operational instructions used for the Codex-assisted draft pass, targeted human refinement, and blind human re-annotation audit. The first two stages are not blind: both had

access to the inherited source-binary label as audit metadata, because they target boundary cases in which contextual support may be over-counted as pair or discarded as non-pair. In contrast, the blind re-annotation audit removes construction metadata; annotators see only dialogue, target, and candidate information.

F.1 Codex-Assisted Draft Pass

Table 22 gives the instruction used for the Codex-assisted draft pass. Codex produces a pair-role proposal and a short evidence-grounded reason for each reviewed candidate. The source-binary label is used only to identify the boundary being audited, not as a gold role label.

F.2 Targeted Human Refinement Rules

Table 23 reports the rules used for targeted human refinement. This stage checks Codex-proposed labels and boundary-sensitive rows; it is not an independent blind relabeling of the full candidate space. Reviewers may inspect the source-binary label, the Codex proposal, and the dialogue context in order to decide whether the proposed boundary movement is valid.

The rules are conservative about `emo-context`. A reviewer confirms `emo-context` only when the candidate supplies observable, non-triggering evidence that makes the target emotion easier to interpret. Otherwise, the reviewer keeps or restores `emo-cause` for direct triggers and `non-pair` for unsupported candidates.

F.3 Blind Human Re-Annotation Audit Instructions

The blind re-annotation audit uses the same role definitions but removes construction metadata from the annotator view. Table 24 gives the instruction. This separation matters because Codex-assisted drafting and targeted human refinement are construction and verification steps, while the blind audit measures whether the pair-role boundary can be recovered without seeing source labels, proposals, final labels, provenance, or sampling strata.

G Responsible Research Statement

G.1 Intended Use and Artifact Release

IEMO-MECP is intended as a research diagnostic artifact for studying label-space compression in ECPE. It supports dataset auditing, benchmark analysis, and evaluation diagnostics, and is not intended for deployed user-facing emotion explanation, clinical or legal decision making, employment decisions, educational assessment, or other individual-level decisions.

The artifact builds on IEMOCAP-derived ConvECPE resources (Busso et al., 2008; Li et al., 2023) and preserves the inherited candidate-pair structure while adding pair-role labels and diagnostic metadata. We will release the IEMO-MECP pair-role labels, split files, candidate-pair identifiers, diagnostic metadata, and code/scripts. We will not redistribute restricted raw audio, video, dialogue content, or other underlying data not created by us.

G.2 Licensing and Data Protection

The pair-role labels, split files, and diagnostic metadata created by us will be released under the Creative Commons Attribution 4.0 International license (CC BY 4.0). The accompanying code and scripts will be released under the MIT License. Candidate-pair identifiers are included only to align the released labels with the inherited candidate-pair structure.

To reduce data-protection risks, the released files are limited to pair-role labels, candidate-pair identifiers, split information, and diagnostic metadata needed to reproduce the analyses. We do not add speaker names, demographic attributes, contact information, or other personally identifying fields. Before release, we check the annotation and metadata files to ensure that no new personally identifying fields are introduced.

G.3 Human Annotation and Ethics Review

Targeted human judging and the blind re-annotation audit were conducted by trained members of the author team, not by crowdworkers or external participants. The annotators were selected internally for their familiarity with NLP annotation, ECPE, and the pair-role definitions, and were not paid per item. The annotation work was part of normal research activity. The author-annotators were informed that their judgments would be used for research validation, dataset construction, and the released IEMO-MECP annotation files. We did not collect personal data from annotators beyond their annotation decisions.

Table 22: Instruction used for the Codex-assisted draft pass over inherited ECPE candidate pairs.

Prompt: Codex-Assisted Draft Pass

```
You are auditing ECPE candidate pairs.

Goal:
Propose one pair role for each reviewed candidate and identify contextual support that may be compressed by the source-binary
      pair/non-pair label space.

You may see the source-binary label. Use it as audit metadata only:
- source pair: the candidate may be a direct cause, or it may be context that was over-counted as causal.
- source non-pair: the candidate may be unsupported, or it may be contextual support missed by the binary label.
Do not treat the source-binary label as the final pair role.

Available information:
- dialogue identifier and split
- target turn, target speaker, target emotion, target text
- candidate turn, candidate speaker, candidate text
- source-binary label or current draft role, when available
- dialogue context

Choose one pair role:

emo-cause:
The candidate directly triggers, explains, states, realizes, or appraises the target emotion. A self-pair is emo-cause only
     when the target utterance itself contains the event, reason, appraisal, or desire-obstacle conflict.

emo-context:
The candidate is not the direct trigger, but it provides interpretive support for the target emotion. It may provide
     background, conflict setup, relationship state, discourse bridge, prior speaker state, situational constraint, or an
     alternative plausible cause. Use emo-context only when removing the candidate would weaken interpretation of the target
     emotion.

non-pair:
The candidate does not provide direct-cause evidence or contextual support. Same-topic continuation, nearby position, speaker
      overlap, lexical overlap, or vague mood match is not enough.

Priority rules:
1. Direct cause beats context.
2. Context is not a weak-cause or uncertain-cause label.
3. For source pair -> emo-context, explain why the candidate supports interpretation but does not directly trigger the target
      emotion.
4. For source non-pair -> emo-context, explain what discourse evidence the binary label missed.
5. If evidence is ambiguous or only topical, choose non-pair.

Output one reviewed row per candidate with:
- pair identifiers
- source-binary label or current draft role, when available
- proposed pair role
- short evidence-grounded reason
```

Table 23: Instruction used for targeted human refinement of Codex-proposed pair roles.

Prompt: Targeted Human Refinement

```
You are validating targeted Codex pair-role proposals.

This is not a blind task. You may see:
- source-binary label
- Codex proposed role and rationale
- transition type
- target and candidate text
- full dialogue context

Task:
Approve, reject, or correct the proposed pair role by assigning one final pair role.

Allowed final pair roles:
- emo-cause
- emo-context
- non-pair

Rules:
1. Treat the Codex label as a proposal, not as authority.
2. Treat the source-binary label as boundary metadata, not as authority.
3. Use emo-cause only for observable direct-cause evidence.
4. Use emo-context only for non-triggering discourse evidence that helps interpret the target emotion.
5. Keep a source pair as emo-cause when it directly explains the target; change it to emo-context only when it is explanatory
      background rather than the trigger.
6. Change a source non-pair to emo-context only when the candidate provides concrete interpretive support missed by the
     binary label.
7. Same-topic relation, adjacency, lexical overlap, or same speaker is insufficient for emo-context.
8. For self-pairs, use emo-cause only when the target utterance contains its own event, reason, appraisal, or desire-obstacle
      conflict.
9. When evidence is too weak or ambiguous to support cause or context, choose non-pair.

Output:
- final pair role
- short evidence-grounded reason
```

Table 24: Instruction used for the blind human re-annotation audit.

Prompt: Blind Pair-Role Re-Annotation

```
You are independently re-annotating candidate-to-target emotion pairs.
This is a blind audit.

You may see:
- dialogue context
- target turn, target speaker, target emotion, target text
- candidate turn, candidate distance, candidate speaker,
  candidate emotion, candidate text

You must not see or use:
- source-binary pair/non-pair label
- Codex proposed role or rationale
- final pair-role label
- label source or provenance
- packet type, sampling stratum, or boundary category
- other annotators' judgments

For each candidate, choose one pair role:

emo-cause:
The candidate directly triggers, explains, states, realizes, or appraises the target emotion.

emo-context:
The candidate is not the direct trigger, but it provides background, setup, relationship state, prior state, discourse bridge,
       or situational information needed to interpret the target emotion.

non-pair:
The candidate does not provide enough evidence for direct causality or contextual support.

Decision rules:
1. Direct cause beats context.
2. Context must be evidence-bearing; proximity or topical similarity is not enough.
3. If the candidate only makes the dialogue generally coherent but does not help interpret the target emotion, choose non-
      pair.
4. If unsure between cause and context, choose the best role and mark causal-chain ambiguity.
5. If unsure because evidence is absent or too weak, choose non-pair.

Output:
- pair role: emo-cause | emo-context | non-pair
- confidence: high | medium | low
- causal-chain ambiguity: yes | no
- optional note for difficult cases
```