LEARNING WHAT TO REJECT: COUNTERFACTUAL PREFERENCE ALIGNMENT FOR EMOTION–CAUSE PAIR EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion–cause pair extraction (ECPE) identifies emotion clauses together with the clauses that explain them. A persistent challenge is that ECPE benchmarks are strongly locality-skewed: annotated causes often occur near their emotions, making relative position a highly predictive shortcut for cause selection. Models can therefore favor nearby plausible clauses without learning the distinction that determines the annotation-supported cause. Existing methods mainly strengthen candidate representation or reasoning, but standard positive supervision provides little targeted contrast against shortcut-compatible alternatives. We introduce CPA-DPO, a counterfactual preference-alignment framework that turns such alternatives into explicit competitors. CPA-DPO constructs reliable, challenging, and controlled preferences: verifier-defined eligibility filters counterfactuals, hardest-valid selection chooses the most competitive eligible alternative, and position-balanced formation prevents position from deterministically revealing the preference label. Standard DPO then optimizes these task-specific comparisons. On the matched Qwen2.5-32B setting, CPA-DPO improves ECPE-CN pair F1 from 83.62 to 85.43 over Vanilla-DPO ($+1.81$). Its advantage grows to $+5.04$ F1 on LONG pairs and $+5.30$ on Rebalanced-CN, while independent hard-negative accuracy rises from 87.92 to 91.35.

1 INTRODUCTION

Emotion–cause pair extraction (ECPE) aims to identify which clauses express emotions and which clauses explain them Xia & Ding (2019). A persistent difficulty is that ECPE benchmarks exhibit a strong locality regularity: annotated causes often occur in the same clause as, or close to, their emotions Ding et al. (2020a); Su et al. (2024). Relative position is therefore highly predictive, allowing a model to perform well by favoring nearby clauses without fully distinguishing an annotation-supported cause from other plausible events. This weakness becomes especially visible for long-distance pairs and when the original position distribution is perturbed. The central challenge is to make locality insufficient for deciding which plausible candidate should win.

Existing ECPE methods mainly strengthen how emotion–cause candidates are represented, matched, or reasoned over. Representative approaches include structured and graph-based interaction Ding et al. (2020a); Chen et al. (2020); Fan et al. (2020); Zhang et al. (2025); Wang et al. (2025), MRC-style extraction and contrastive reasoning Zhou et al. (2022); Cheng et al. (2023); Mai et al. (2024), and prompt- or LLM-based reasoning Gu et al. (2024); Yu et al. (2025). These advances improve candidate comparison, but the supervision itself typically provides little targeted contrast against nearby, emotionally salient, or topically compatible alternatives. The model is told which cause is annotation-supported, while plausible shortcut-compatible mistakes are rarely made explicit competitors. This motivates our central question: *what supervision makes the cause selector distinguish the annotated cause from alternatives that remain compatible with the benchmark's shortcut cues?*

Figure 1 illustrates the resulting preference-construction problem. Useful chosen–rejected comparisons should satisfy three requirements. They should be *reliable*, so that the rejected alternative provides a clear preference direction; *challenging*, so that coarse relevance or locality is insufficient to separate it from the gold cause; and *controlled*, so that systematic construction cues do not reveal

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

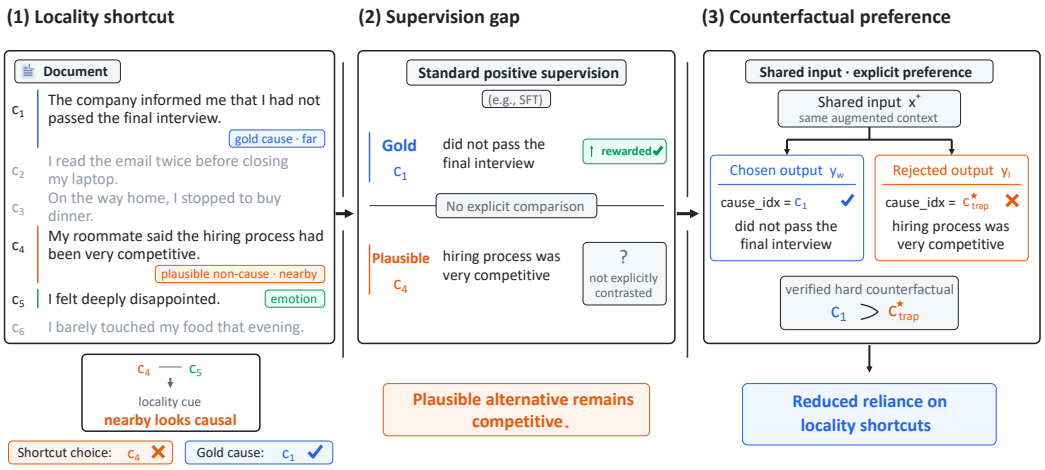


Figure 1: **From locality shortcuts to counterfactual preference supervision.** Locality can make nearby plausible non-causes competitive. CPA-DPO addresses this supervision gap by explicitly preferring the annotation-supported cause over a verified hard counterfactual.

the preference label. These requirements make preference construction, rather than the preference objective itself, the central technical problem.

We introduce CPA-DPO (Counterfactual Preference Alignment with DPO), a cause-targeted framework built around these three requirements. A teacher first generates context-preserving counterfactual alternatives. A frozen document-level verifier defines the reliable region by retaining candidates that sufficiently reduce support for the target emotion–cause relation. Within this region, *hardest-valid* selection chooses the candidate with the highest residual support, yielding a competitive rejected alternative. Finally, position-balanced preference formation swaps the semantic roles of the gold cause and selected counterfactual across original and appended positions, so that position does not deterministically reveal the preference label. The resulting chosen and rejected outputs share the same augmented input, while semantic role determines which output is preferred.

Standard DPO then optimizes these task-specific preferences. Eligibility determines which comparisons are trustworthy, hardest-valid selection determines which trustworthy comparison is most informative, and position-balanced formation controls construction-label leakage. Together, these components make plausible shortcut-compatible alternatives explicitly lose against annotation-supported causes. All preference construction is performed offline; inference retains the original two-stage ECPE pipeline without teacher generation or verifier scoring.

The design reflects a broader supervision principle: when shortcut features remain predictive in the training distribution, preference learning should compare outputs for which those features are insufficient to determine the winner. CPA-DPO operationalizes this principle by constructing counterfactual comparisons that directly target the distinction obscured by locality.

Our experiments test both effectiveness and mechanism. On the matched Qwen2.5-32B setting, CPA-DPO improves ECPE-CN pair F1 from 83.62 to 85.43 over Vanilla-DPO ($+1.81$). Its advantage grows to $+5.04$ F1 on LONG pairs and $+5.30$ on Rebalanced-CN, while worst-case degradation under distractor relocation decreases from 4.27 to 2.72. Independent hard-negative accuracy also increases from 87.92 to 91.35 ($+3.43$), showing that the learned discrimination transfers beyond the counterfactual candidates used during training. Construction-metadata probes remain near chance, and component ablations separate the contributions of eligibility and hardest-valid selection.

Our contributions are threefold. **First**, we identify a supervision gap behind locality shortcuts in ECPE: annotation-supported causes are rarely contrasted against targeted shortcut-compatible alternatives. **Second**, we introduce CPA-DPO, which constructs reliable, challenging, and controlled chosen–rejected supervision through verifier-defined eligibility, hardest-valid counterfactual selection, and position-balanced preference formation. **Third**, we show that the benefit of this supervision grows as locality becomes less informative and transfers to independently constructed hard negatives, with leakage probes and ablations supporting the intended construction mechanism.

2 Related Work

Our work connects three lines of research: ECPE modeling and positional bias, counterfactual approaches to shortcut learning, and preference optimization. Together, they motivate the use of task-specific preference construction to make shortcut-compatible errors informative training signals.

2.1 Emotion–Cause Pair Extraction and Positional Bias

Emotion cause extraction predicts causes for given emotions Lee et al. (2010); Gui et al. (2016), while ECPE jointly identifies emotion–cause pairs from documents Xia & Ding (2019). Subsequent work has improved extraction through structured prediction, inter-clause interaction, graph and hierarchical modeling, and MRC- or LLM-based reasoning Ding et al. (2020a); Fan et al. (2020); Cheng et al. (2020); Liu et al. (2022); Su et al. (2024); Zhou et al. (2022); Cheng et al. (2023); Gu et al. (2024); Yu et al. (2025). These methods primarily strengthen how candidate clauses are represented, compared, or searched. Position remains an important regularity in ECPE: DECC, for example, shows substantial performance changes when the original position distribution is rebalanced Wu et al. (2024). CPA-DPO targets the supervision used to distinguish an annotation-supported cause from a plausible competing clause.

2.2 Shortcut Learning and Counterfactual Supervision

Shortcut learning occurs when models exploit predictive regularities that work on the training distribution without capturing the intended task distinction. Counterfactually augmented data, contrast sets, and invariance-oriented analyses expose such behavior by altering superficial cues while preserving task-relevant structure Kaushik et al. (2020); Gardner et al. (2020); Feder et al. (2022); Veitch et al. (2021). This perspective is especially relevant to ECPE, where nearby or semantically compatible clauses can remain plausible non-causes. Counterfactual learning has also been explored within ECPE, including counterfactual contrastive MRC for improving discriminative representations Mai et al. (2024). CPA-DPO uses counterfactuals to create explicit chosen–rejected supervision between annotation-supported causes and plausible competing alternatives.

2.3 Preference Learning and Task-Specific Contrast Construction

Preference learning trains models from relative judgments between outputs. RLHF and Direct Preference Optimization (DPO) established chosen–rejected supervision as an effective alignment paradigm Christiano et al. (2017); Ouyang et al. (2022); Rafailov et al. (2023). For structured extraction, the usefulness of this signal depends on how the comparison is constructed: easy negatives provide weak discrimination, ambiguous alternatives weaken preference reliability, and systematic differences between chosen and rejected outputs can themselves become label cues. Preference construction is therefore part of the task design, including the control of construction-label leakage.

CPA-DPO addresses this issue by constructing preferences that are reliable, challenging, and controlled, with standard DPO used as the optimization layer. This positions CPA-DPO as a supervision-level approach to ECPE shortcut mitigation: plausible shortcut-compatible alternatives are made explicit competitors during training, while the DPO objective remains standard. The primary contribution is the construction of reliable, challenging, and controlled preferences rather than a modification of DPO itself.

3 Methodology

CPA-DPO constructs task-specific preference supervision for emotion-conditioned cause selection. It combines context-preserving counterfactual generation, verifier-defined eligibility, hardest-valid selection, and position-balanced preference formation to produce reliable, challenging, and controlled chosen–rejected comparisons, which are optimized with standard DPO. Figure 2 summarizes the pipeline.

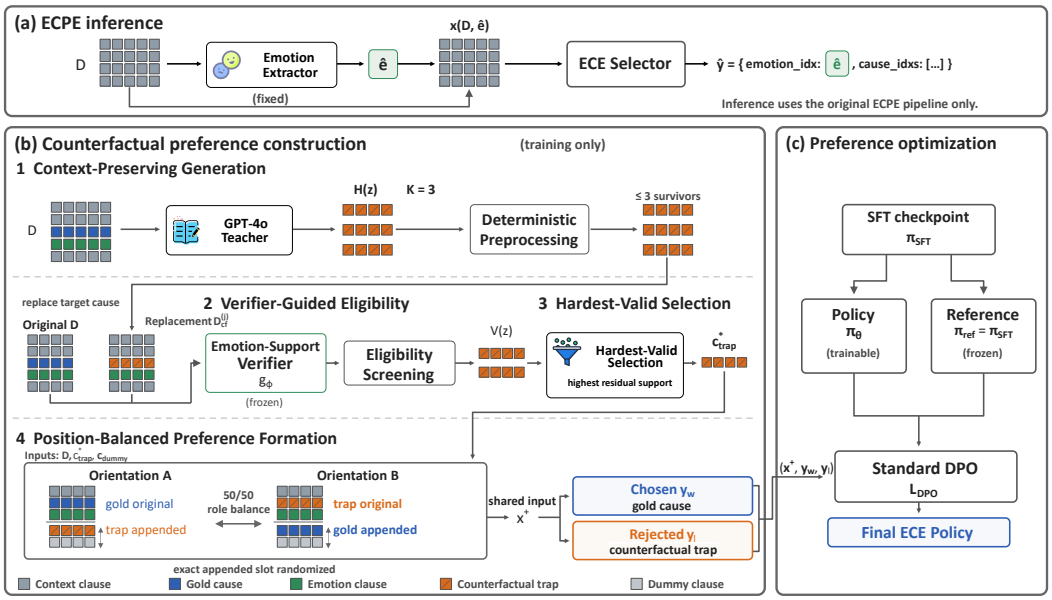


Figure 2: **Overview of CPA-DPO.** (a) Inference retains the original ECPE pipeline. (b) During training, a teacher generates counterfactual alternatives, a frozen verifier defines eligible candidates, hardest-valid selection chooses the strongest eligible alternative, and position-balanced formation constructs the preference pair. (c) Standard DPO optimizes the resulting preferences from the SFT initialization and frozen reference.

3.1 Task and Learning Setup

Let $D = (c_1, \ldots, c_n)$ be a document and e a target emotion clause. Following the standard two-stage ECPE setting, the emotion-conditioned cause extractor (ECE) predicts

$$y = \{\texttt{emotion_idx} : e,\ \texttt{cause_idxs} : [k_1, \ldots, k_m]\}. \tag{1}$$

CPA-DPO changes the supervision for this second-stage cause selector while keeping the upstream emotion extractor and inference interface fixed.

Preference construction operates on

$$z = (D, e, k_{\text{gold}}, c_{\text{gold}}), \qquad c_{\text{gold}} = c_{k_{\text{gold}}}, \tag{2}$$

where k_{gold} denotes one annotation-supported cause. Let $\mathcal{G}(z)$ be the annotated cause-index list serialized in the chosen completion, with $k_{\text{gold}} \in \mathcal{G}(z)$. For single-cause records, $\mathcal{G}(z) = [k_{\text{gold}}]$; for multi-cause-derived records, it may retain additional annotated causes. Further data details are given in Appendix D.3.

3.2 Counterfactual Preference Construction

Context-preserving generation. Easy negatives provide little discrimination, so a teacher generates $K = 3$ context-preserving alternatives:

$$\mathcal{H}(z) = \{c_{\text{trap}}^{(1)}, \ldots, c_{\text{trap}}^{(K)}\}. \tag{3}$$

Deterministic preprocessing removes malformed, duplicate, and exact-source outputs, yielding $\mathcal{H}_{\text{rule}}(z)$. Prompt and preprocessing details are given in Appendices B.2 and B.1.

Verifier-guided eligibility. A frozen document-level verifier measures target-emotion support:

$$S(D, e) = P_{g_\phi}(\text{Yes} \mid D, e), \qquad A_{\text{orig}} = S(D, e). \tag{4}$$

For each candidate $c_{\text{trap}}^{(j)}$, we replace the targeted gold cause at its original position:

$$D_{\text{cf}}^{(j)} = D_{c_{\text{gold}} \to c_{\text{trap}}^{(j)}}, \qquad A_{\text{cf}}^{(j)} = S(D_{\text{cf}}^{(j)}, e), \tag{5}$$

and compute the support reduction

$$\Delta_j = A_{\text{orig}} - A_{\text{cf}}^{(j)}. \tag{6}$$

Eligible candidates satisfy

$$\mathcal{V}(z) = \left\{ c_{\text{trap}}^{(j)} \in \mathcal{H}_{\text{rule}}(z) : A_{\text{orig}} \geq 0.8,\ \Delta_j \geq 0.3 \right\}. \tag{7}$$

Records with $\mathcal{V}(z) = \varnothing$ produce no preference pair.

Hardest-valid selection. Among eligible candidates, we select the one with the highest residual target-emotion support:

$$c_{\text{trap}}^{\star} = \arg \max_{c_{\text{trap}}^{(j)} \in \mathcal{V}(z)} A_{\text{cf}}^{(j)}. \tag{8}$$

Eligibility controls reliability, while hardest-valid selection chooses the most competitive candidate within that reliable region. Verifier and selection details are provided in Appendices B.3–B.5.

Position-balanced preference formation. To prevent original/appended position from deterministically revealing the preference label, we sample $b \sim \text{Bernoulli}(0.5)$ to choose whether the gold cause or $c_{\text{trap}}^{\star}$ occupies the original target-cause position. Independently, $\rho \sim \text{Bernoulli}(0.5)$ randomizes the two appended slots.

The appended clauses are

$$(a_{n+1}, a_{n+2}) = \begin{cases} \text{Perm}_{\rho}(c_{\text{trap}}^{\star}, c_{\text{dummy}}), & b = 0, \\ \text{Perm}_{\rho}(c_{\text{gold}}, c_{\text{dummy}}), & b = 1, \end{cases} \tag{9}$$

and the augmented document is

$$\widetilde{D}_b = \begin{cases} \text{Append}(D, a_{n+1}, a_{n+2}), & b = 0, \\ \text{Append}(D_{c_{\text{gold}} \to c_{\text{trap}}^{\star}}, a_{n+1}, a_{n+2}), & b = 1. \end{cases} \tag{10}$$

Thus, for $b = 0$ the gold cause remains at its original position and the trap is appended; for $b = 1$ the trap occupies the original position and the gold cause is appended. Let k_{gold}^{+} and k_{trap}^{+} denote their positions in $\widetilde{D}_b$.

Let $\mathcal{G}_b^{+}(z)$ be $\mathcal{G}(z)$ after remapping the relocated gold index when necessary. Both outputs share

$$x^{+} = x(\widetilde{D}_b, e), \tag{11}$$

and form the preference triple

$$y_w = \mathcal{G}_b^{+}(z), \qquad y_l = [k_{\text{trap}}^{+}], \qquad (x^{+}, y_w, y_l). \tag{12}$$

For single-cause records, $\mathcal{G}_b^{+}(z) = [k_{\text{gold}}^{+}]$. The two outputs therefore share the same augmented input and structured output schema, while semantic role determines the preference label. Section 4.4 evaluates the position balance with leakage probes; structural checks are given in Appendices B.6 and B.7.

3.3 Preference Optimization and Inference

The SFT policy initializes the trainable model and serves as the frozen DPO reference:

$$\pi_{\text{ref}} = \pi_{\text{SFT}}. \tag{13}$$

Standard DPO Rafailov et al. (2023) optimizes the preference triples in Eq. equation 12; complete settings are given in Appendix C.2. All headline results use the raw post-DPO policy.

At inference, the frozen emotion extractor predicts target emotions and the final ECE selector generates Eq. equation 1. Preference construction is entirely offline, so deployment retains the same two-stage ECPE inference path as the matched baselines.

Table 1: **Main results.** (a) Published benchmark context; "–" denotes no verified comparable result in the cited literature. (b) Qwen2.5-32B comparison. SFT, Vanilla-DPO, and Balanced CPA-DPO form the matched trainable comparison, while 4-shot Prompt and DECC-style 4-shot provide prompt-based reference points. Gold-E isolates cause selection; Worst Drop is the worst relocation degradation (lower is better); Indep. HN is Independent HN Accuracy.

(a) Published benchmark context

Method	Backbone / Setting	ECPE-CN F1	ECPE-EN F1	Rebalanced-CN F1
ECPE-2D (Ding et al., 2020a)	–	68.89	50.73	41.73
RANKCP (Wei et al., 2020)	BERT+GCN	73.60	44.63	41.22
ECPE-MLL (Ding et al., 2020b)	–	74.52	51.21	45.57
A2Net (Chen et al., 2022)	BERT	76.34	–	–
MM-R (Zhou et al., 2022)	MRC-based PLM	80.62	52.08	–
HSE-TTLS (Su et al., 2024)	BERT	82.18	53.08	–
CokECPE (Hu & Xie, 2025)	BERT	77.79	–	–
BMECPE (Li et al., 2026)	Chinese-RoBERTa-wwm	85.12	62.71*	–
GPT-3.5-turbo DECC (4-shot) (Wu et al., 2024)†	GPT-3.5-turbo / 4-shot	69.95	50.35	61.38

(b) Qwen2.5-32B comparison

Method	Gold-E F1	ECPE-CN F1	ECPE-EN F1	Rebalanced-CN F1	Worst Drop ↓	Indep. HN
4-shot Prompt	–	64.20	54.10	55.60	–	–
DECC-style 4-shot	–	72.10	60.80	65.70	–	–
SFT	88.46	73.07	66.02	61.80	6.16	79.85
Vanilla-DPO	89.25	83.62	67.15	74.65	4.27	87.92
Balanced CPA-DPO (Ours)	**90.18**	**85.43**	**68.12**	**79.95**	**2.72**	**91.35**

*The BMECPE English result is reported on its NTCIR-13 benchmark under an 80/10/10 train/validation/test protocol and is included only as literature context because its dataset configuration is not directly matched to our ECPE-EN setting. †DECC results are taken from the original GPT-3.5-turbo 4-shot evaluation, which uses human semantic-equivalence judgments for generative predictions.

4 EXPERIMENTS

We evaluate CPA-DPO through matched performance, shortcut-sensitive tests, cause discrimination, construction validity, and mechanism analysis.

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate primarily on ECPE-CN under the standard document-level ten-fold protocol Xia & Ding (2019); Ding et al. (2020a); Fan et al. (2020); Su et al. (2024); Hu & Xie (2025); Li et al. (2026), using end-to-end pair F1 as the primary metric. ECPE-EN provides cross-dataset evaluation. Shortcut-sensitive tests use Rebalanced-CN, LONG pairs ($|e - k| \geq 2$), benign appending, and distractor relocation. Gold-E ECE isolates cause selection, while constructed and independent hard negatives probe cause/non-cause discrimination. Full definitions are in Appendix A.

Matched protocol. Our controlled comparison keeps the Qwen2.5-32B backbone, SFT initialization, frozen emotion extractor, fold-wise upstream predictions, input/output interface, DPO reference, optimization configuration, and direct inference procedure fixed across Vanilla-DPO and CPA-DPO. We additionally include a direct 4-shot baseline, a matched DECC-style adaptation (Wu et al., 2024), and Random-DPO for hard-negative analysis. Construction, training, and protocol details are in Appendices D.3, C.2, and A.2.

4.2 MAIN RESULTS

Table 1 reports published benchmark context and the matched Qwen2.5-32B comparison.

Under the matched trainable setting, Balanced CPA-DPO improves ECPE-CN F1 from 83.62 to 85.43 over Vanilla-DPO, a $+1.81$-point gain. We next examine the stronger pattern under shortcut-sensitive evaluation.

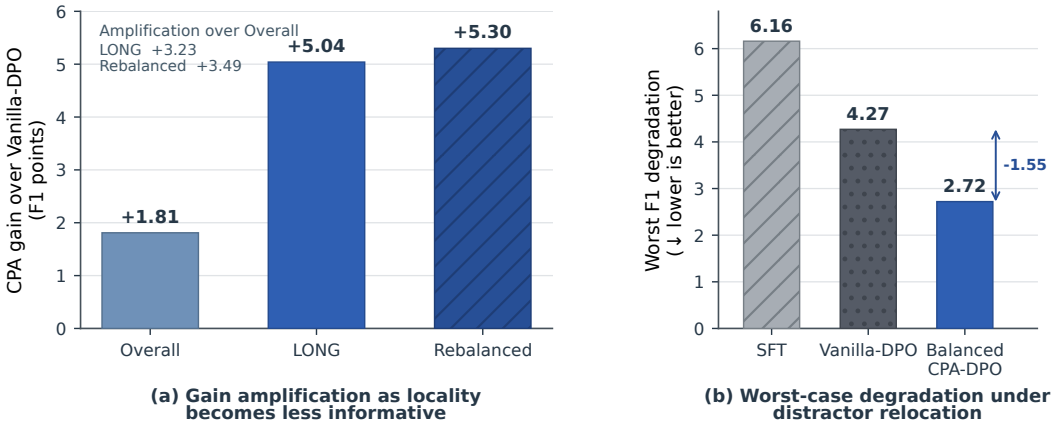


Figure 3: **Shortcut-sensitive evaluation on ECPE-CN.** (a) Balanced CPA-DPO improves Vanilla-DPO by $+1.81$ F1 overall, while the gains increase to $+5.04$ on LONG pairs and $+5.30$ under position-rebalanced evaluation. Relative to the Overall gain, the corresponding amplification is $+3.23$ and $+3.49$ F1 points. (b) Worst-case degradation under distractor relocation decreases from 4.27 for Vanilla-DPO to 2.72 for Balanced CPA-DPO, while remaining non-zero. Lower is better in panel (b).

Table 2: **Cause/non-cause discrimination.** Independent HN uses separately constructed hard negatives and therefore tests transfer beyond the exact counterfactual preferences used during training. Higher is better.

Method	Indep. HN Acc.
SFT	79.85
Random-DPO	84.10
Vanilla-DPO	87.92
Balanced CPA-DPO	**91.35**

Table 3: **Construction-label leakage diagnostics for Balanced CPA-DPO.**

Diagnostic	Balanced CPA
$P(\text{appended} \mid \text{chosen})$	50.2%
$P(\text{appended} \mid \text{rejected})$	49.8%
Position-only probe accuracy	50.6%
Full metadata probe accuracy	52.1%
Chance baseline	50.0%

4.3 Shortcut Sensitivity and Cause Discrimination

Balanced CPA-DPO exhibits a markedly larger advantage when locality becomes less informative. Its gain over Vanilla-DPO increases from $+1.81$ F1 overall to $+5.04$ on LONG pairs and $+5.30$ on Rebalanced-CN.

We quantify fold-wise amplification as

$$A_f = \left(F^{\text{CPA}}_{\text{LONG},f} - F^{\text{Vanilla}}_{\text{LONG},f}\right) - \left(F^{\text{CPA}}_{\text{Overall},f} - F^{\text{Vanilla}}_{\text{Overall},f}\right), \tag{14}$$

and analogously define

$$B_f = \left(F^{\text{CPA}}_{\text{Rebalanced},f} - F^{\text{Vanilla}}_{\text{Rebalanced},f}\right) - \left(F^{\text{CPA}}_{\text{Overall},f} - F^{\text{Vanilla}}_{\text{Overall},f}\right). \tag{15}$$

The mean LONG-minus-Overall amplification is $+3.23$ F1, while the Rebalanced-minus-Overall amplification is $+3.49$ F1.

Worst-case degradation under distractor relocation decreases from 4.27 to 2.72; complete shortcut-sensitive values are reported in Appendix D.1.

Independent HN accuracy rises from 87.92 to 91.35 ($+3.43$). This gain indicates that improved cause/non-cause discrimination extends beyond the exact counterfactual candidates used during preference construction.

4.4 Construction Validity and Mechanism Analysis

We first test construction-label leakage. Balanced formation nearly equalizes appended-position frequency between chosen and rejected candidates (50.2% versus 49.8%). The position-only and

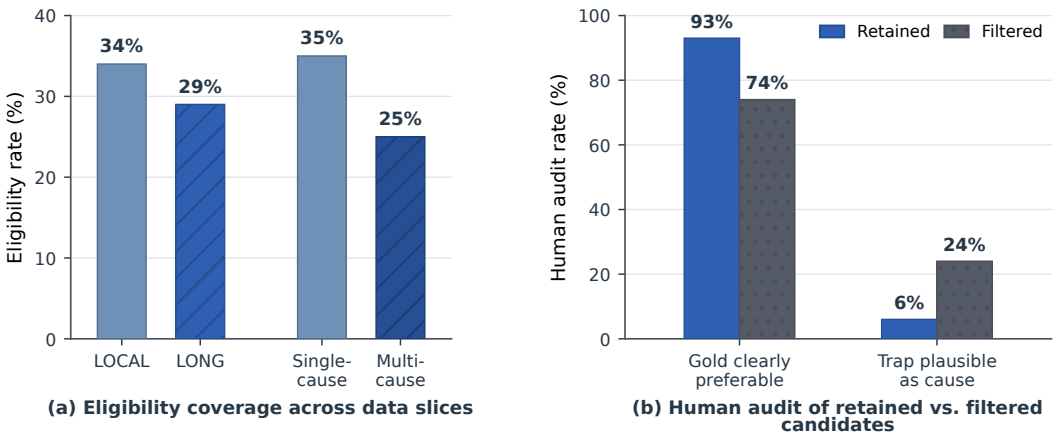


Figure 4: **Eligibility coverage and human validation of verifier-based filtering.** (a) Eligibility rate across distance and cause-multiplicity slices, defined as the fraction of source records for which the verifier-defined eligible set $V(z)$ is non-empty. (b) Human audit of retained versus filtered candidates.

full-metadata probes reach 50.6% and 52.1% accuracy, respectively, close to the 50.0% chance baseline.

Figure 4(a) shows eligibility rates of 34% for LOCAL, 29% for LONG, 35% for single-cause, and 25% for multi-cause records, indicating that verifier-defined eligibility remains viable across the examined slices. Figure 4(b) provides complementary human evidence: the gold cause is clearly preferable in 93% of retained versus 74% of filtered comparisons, while the trap is judged plausible as a cause in 6% versus 24%, respectively. These results support eligibility as a reliability filter. Full construction diagnostics are provided in Appendix D.3 and Appendix D.4.

We then ablate eligibility and hardest-valid selection. Position balancing is evaluated through the leakage probes above because its role is construction control.

Table 4: **Mechanism ablation on ECPE-CN.** Random-Valid isolates hardest-valid selection within the eligible set; HN denotes Independent HN Accuracy. Higher is better.

Variant	Overall F1	LONG F1	Indep. HN
Random negative	83.77	70.91	88.31
No eligibility	84.09	72.02	89.33
Random-valid	84.48	73.41	90.24
Hardest-valid Balanced CPA-DPO	**85.43**	**75.46**	**91.35**

Eligibility improves Random-Valid over the unfiltered counterfactual variant by $+0.39$ Overall, $+1.39$ LONG, and $+0.91$ Independent-HN points. Hardest-valid selection provides a further $+0.95$ Overall, $+2.05$ LONG, and $+1.11$ HN improvement over random selection within the eligible region. The larger differences on LONG and HN are consistent with these components being most useful when coarse shortcut cues are insufficient.

The same balanced preference data reach 84.45 Overall / 73.82 LONG / 90.31 HN with a pairwise ranking objective, while DPO reaches 85.43 / 75.46 / 91.35, corresponding to gains of $+0.98$ Overall, $+1.64$ LONG, and $+1.04$ HN. Full objective-control and threshold analyses are reported in Appendices D.5 and D.2.

4.5 Operating Boundaries

Balanced CPA-DPO introduces additional offline computation for teacher generation and verifier scoring, while deployment retains the same direct two-stage ECPE path as the matched baselines. Worst-case relocation incurs a 2.72-point F1 drop. Further scope and limitations are discussed in Sec. 5.

5 Discussion and Limitations

Supervision-level effect. CPA-DPO frames shortcut-sensitive ECPE as a supervision-design problem. Instead of relying only on stronger representations or reasoning, it makes plausible shortcut-compatible alternatives explicit competitors during training. Verifier-defined eligibility, hardest-valid selection, and position-balanced formation respectively control preference reliability, difficulty, and construction leakage. The larger gains on LONG and Rebalanced-CN, together with hard-negative and leakage analyses, indicate that targeted preference construction is effective when locality cues become less informative. This supervision strategy is complementary to stronger ECPE encoders, structured models, and discourse reasoning.

Scope and limitations. The current construction depends on the quality of teacher-generated counterfactuals and verifier scoring, motivating future human-audited or multi-verifier construction. It is also cause-targeted: each comparison centers on one annotation-supported cause and one counterfactual alternative, while some multi-cause records retain additional annotated causes in the chosen output. Extending the framework to explicitly set-valued preferences is therefore a natural next step. Preference construction adds offline generation and verification cost, but deployment retains the same direct two-stage inference path as the matched baselines. Broader multilingual, cross-domain, and long-context evaluation is another promising direction.

6 Conclusion

We introduced CPA-DPO, a counterfactual preference-alignment framework for reducing reliance on locality shortcuts in emotion–cause pair extraction. CPA-DPO turns plausible shortcut-compatible errors into explicit chosen–rejected supervision through verifier-defined eligibility, hardest-valid selection, and position-balanced preference formation. Across matched, long-distance, position-rebalanced, and perturbation evaluations, a consistent pattern emerges: the advantage of CPA-DPO becomes substantially larger as locality cues become less predictive. Complementary evidence from independent hard-negative discrimination, mechanism ablations, and near-chance construction-metadata probes further supports the contribution of task-specific preference construction. The matched gain over Vanilla-DPO is $+1.81$ F1 overall and increases to $+5.04$ on LONG and $+5.30$ on Rebalanced-CN, while Independent HN Accuracy improves by $+3.43$. These results support preference construction as the primary contribution, with standard DPO retained as the optimization objective.

More broadly, our results suggest that supervision design can provide a practical route to reducing shortcut reliance. When superficial cues remain predictive in the training distribution, constructing comparisons in which those cues are insufficient to determine the preferred output can provide more targeted learning signals. This perspective may extend beyond ECPE: counterfactual preference construction offers a potentially broader supervision principle for making shortcut-compatible mistakes explicitly lose while preserving the task-relevant distinction the model should learn.

AI Use Statement

Generative-AI tools were used during manuscript preparation for language editing, structural revision, consistency checking, LaTeX assistance, and figure-design support. The authors reviewed and edited all AI-assisted content, verified the reported numerical results and methodological descriptions against the underlying experiment artifacts, and take full responsibility for the final manuscript and its claims. The use of GPT-4o as a counterfactual generator is part of the proposed experimental methodology and is separately documented in the method and appendix.

Ethics Statement

This work studies emotion–cause pair extraction using existing research benchmarks and does not introduce a new collection of personal data. The proposed training procedure uses model-generated counterfactual clauses and a frozen verifier to construct preference supervision; these components may inherit biases or errors from the models on which they depend. The manual audits reported in the appendix evaluate generated clauses and annotation-supported relations for research validation.

The extracted "cause" relations should be interpreted according to the benchmark annotations and task definition. They are not intended as verified psychological, medical, legal, or real-world causal explanations. Accordingly, predictions from the evaluated systems should not be used as standalone evidence for consequential decisions about individuals. The paper reports the scope and remaining limitations of the construction and evaluation procedures to support transparent interpretation of the results.

Reproducibility Statement

The main paper specifies the task formulation, counterfactual preference construction, verifier-defined eligibility, hardest-valid selection, position-balanced preference formation, DPO optimization, and direct inference protocol. Appendix A.2 documents the matched evaluation protocol and fold-wise data isolation. Appendix B provides implementation details for candidate generation, deterministic filtering, verifier scoring, eligibility, hardest-valid selection, position-balanced placement, and preference integrity. Appendix C reports the confirmed training configuration, while Appendix C.5 specifies the test-time decoding and parsing protocol.

All main comparisons report the post-DPO policy directly. Additional reproducibility evidence includes the final preference-data composition in Appendix D.3, threshold sensitivity in Appendix D.2, computational characteristics in Appendix C.4, and fold-level statistical analysis in Appendix D.12. The supplementary material contains the complete prompts, configuration and audit details, extended experimental results, and qualitative diagnostics used to support the reported claims.

References

Shunjie Chen, Xiaochuan Shi, Jingye Li, Shengqiong Wu, Hao Fei, Fei Li, and Donghong Ji. Joint alignment of multi-task feature and label spaces for emotion cause pair extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6955–6965, October 2022.

Xinhong Chen, Qing Li, and Jianping Wang. A unified sequence labeling model for emotion cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 208–218. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.18. URL https://aclanthology.org/2020.coling-main.18/.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Hua Yu, and Qing Gu. A symmetric local search network for emotion-cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 139–149. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.12. URL https://aclanthology.org/2020.coling-main.12/.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Cong Wang, Shiping Ge, and Qing Gu. A consistent dual-MRC framework for emotion-cause pair extraction. *ACM Transactions on Information Systems*, 41(4):105:1–105:27, 2023. doi: 10.1145/3558548.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zixiang Ding, Rui Xia, and Jianfei Yu. 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*, pp. 3161–3170. Association for Computational Linguistics, 2020a. doi: 10.18653/v1/2020.acl-main.288. URL https://aclanthology.org/2020.acl-main.288/.

Zixiang Ding, Rui Xia, and Jianfei Yu. 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*, pp. 3574–3583. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.emnlp-main.290. URL https://aclanthology.org/2020.emnlp-main.290/.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. Transition-based directed graph construction for emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3707–3717. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.342. URL https://aclanthology.org/2020.acl-main.342/.

Amir Feder, Katherine Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10: 1138–1158, 2022.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, Ben Zhou, and Yilun Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020.

Xue Gu, Zhihan Zhou, Ziyao Meng, Jian Li, Tiago Gomes, Adriano Tavares, and Hao Xu. 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*, pp. 5678–5688. ELRA and ICCL, 2024.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, and Yu Zhou. Event-driven emotion cause extraction with corpus construction. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1639–1649. Association for Computational Linguistics, 2016.

Guimin Hu and Jiayuan Xie. Improving representation with intra- and inter-clause commonsense injection for emotion-cause pair extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:1974–1986, 2025. doi: 10.1109/TASLPRO.2025.3525957.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pp. 45–53. Association for Computational Linguistics, 2010.

Guorui Li, Yaxin Wen, Cong Wang, Lihong Cao, Sancheng Peng, and Chao Deng. A method for extracting emotion–cause pairs based on bidirectional machine reading comprehension. *Engineering Applications of Artificial Intelligence*, 166:113584, 2026. doi: 10.1016/j.engappai.2025.113584.

Junlong Liu, Xichen Shang, and Qianli Ma. Pair-based joint encoding with relational graph convolutional networks for emotion-cause pair extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5339–5351. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.358. URL https://aclanthology.org/2022.emnlp-main.358/.

Hanjie Mai, Xuejie Zhang, Jin Wang, and Xiaobing Zhou. A machine reading comprehension model with counterfactual contrastive learning for emotion-cause pair extraction. *Knowledge and Information Systems*, 66(6):3459–3476, 2024. doi: 10.1007/s10115-024-02062-1.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xinxin Su, Zhen Huang, Yixin Su, Bayu Distiawan Trisedya, Yong Dou, and Yunxiang Zhao. Hierarchical shared encoder with task-specific transformer layer selection for emotion-cause pair extraction. *IEEE Transactions on Affective Computing*, 15(4):1934–1948, 2024. doi: 10.1109/TAFFC.2024.3390223.

Victor Veitch, Dhanya Sridhar, and David M. Blei. Counterfactual invariance to spurious correlations in text classification. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16196–16208, 2021.

Yuwei Wang, Yuling Li, Kui Yu, and Jing Yang. A semantic structure-based emotion-guided model for emotion-cause pair extraction. *Pattern Recognition*, 161:111296, 2025. doi: 10.1016/j.patcog.2024.111296.

Penghui Wei, Wenji Ji, Changliang Li, Liang Huang, Xin Wang, and Kewei Tu. 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*, pp. 3171–3181. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.289. URL https://aclanthology.org/2020.acl-main.289/.

Jialiang Wu, Yi Shen, Ziheng Zhang, and Longjun Cai. Enhancing large language model with decomposed reasoning for emotion cause pair extraction. *arXiv preprint arXiv:2401.17716*, 2024.

Rui Xia and Zixiang Ding. 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*, pp. 1003–1012. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-1096/.

Zhaoxin Yu, Xinglin Xiao, and Wenji Mao. 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)*, pp. 10278–10293. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.516.

Ming Zhang, Zhe Chen, Vasile Palade, Tao Lu, Liya Wang, Junchi Zhang, and Yanduo Zhang. Double-graph representation with relational enhancement for emotion–cause pair extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):10859–10873, 2025. doi: 10.1109/TNNLS.2025.3527767.

Changzhi Zhou, Dandan Song, Jing Xu, and Zhijing Wu. A multi-turn machine reading comprehension framework with rethink mechanism for emotion-cause pair extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 6726–6735. International Committee on Computational Linguistics, 2022. URL https://aclanthology.org/2022.coling-1.583/.

A Dataset, Protocol, and Evaluation Audits

This appendix provides additional dataset statistics and protocol-level audits that complement the experimental setup in the main paper. The core empirical evidence—including the standard ECPE results, shortcut-sensitive evaluations, cause/non-cause discrimination, and component ablations—is reported in the main text. Here, we provide the complete emotion–cause distance distribution and verify that the matched SFT, Vanilla-DPO, and Balanced CPA-DPO comparisons do not differ through upstream emotion predictions, backbone capacity, input formatting, decoding configuration, or test-set exposure.

Reporting scope. Unless otherwise stated, all results in this appendix refer to the final position-balanced CPA-DPO construction and the raw post-DPO policy, with no adapter interpolation at inference or checkpoint selection. Supplementary threshold, backbone, RECCON, hard-negative, slice, fold-level, and qualitative analyses below are treated as part of the same final evidence set. We do not claim exact source-ID equality between independently constructed Vanilla-DPO and CPA-DPO preference datasets unless it is directly audited. Where an audit is available only as aggregate percentages, we report it descriptively rather than reconstructing unavailable denominators or inferential statistics.

A.1 ECPE-CN Locality Distribution

The motivation for examining positional shortcut dependence arises from a pronounced locality regularity in ECPE-CN. For an annotated emotion–cause pair (e, k), we define the absolute clause distance as

$$d(e, k) = |e - k|, \tag{16}$$

where e and k denote the clause indices of the emotion and its annotated cause, respectively.

Table 5 reports the complete distance distribution over the 2,167 annotated emotion–cause pairs in ECPE-CN. Of these pairs, 511 (23.58%) occur within the same clause ($|e - k| = 0$), while 1,342 (61.93%) occur at distance 1. Taken together, 1,853 pairs, or 85.51% of the benchmark, occur within one clause of the target emotion. Only 314 pairs (14.49%) satisfy the long-distance criterion used throughout this paper,

$$\text{LONG} : |e - k| \geq 2. \tag{17}$$

The distribution therefore provides a strong statistical regularity: a rule that preferentially selects clauses close to the target emotion is compatible with the large majority of the standard benchmark. This observation does not by itself establish that a trained model actually uses position as its decision rule. Rather, it shows that locality is an available and highly predictive cue that can support shortcut-based prediction.

This distinction motivates the shortcut-sensitive evaluations in the main paper. LONG examples represent the minority regime in which the dominant within-one-clause regularity is less directly useful. Likewise, the position-rebalanced evaluation weakens the predictive value of the original locality distribution, while controlled relocation tests explicitly modify distractor positions without changing the annotated emotion–cause relation.

The quantitative results are consistent with this interpretation. Relative to Vanilla-DPO, Balanced CPA-DPO improves overall ECPE-CN F1 by 1.81 points, whereas the improvement increases to 5.04 points on LONG examples. The corresponding advantage is 5.30 points on the position-rebalanced evaluation. Relative to the Overall gain, the point-estimate amplification is 3.23 points for LONG and 3.49 points for Rebalanced-CN. These enlarged gains when locality becomes less informative are consistent with reduced reliance on positional shortcuts. They should not be interpreted as evidence that positional information has been eliminated or that CPA-DPO is position invariant.

A.2 Matched-Protocol Audit

The primary attribution in our matched experiments compares SFT, Vanilla-DPO, and Balanced CPA-DPO under an otherwise fixed prediction pipeline. We therefore audit the main protocol components that could introduce confounding differences between these systems.

All three matched systems use the same Qwen2.5-32B backbone configuration and the same frozen emotion extractor. Within each fold, emotion predictions are generated once and reused unchanged by the SFT, Vanilla-DPO, and Balanced CPA-DPO cause-selection systems. The emotion-conditioned input template, structured output format, direct evaluator, and decoding configuration are also identical across the matched comparison.

Consequently, the compared systems differ in the supervision used to train the emotion-conditioned cause selector rather than in upstream emotion predictions, model capacity, or test-time interface. SFT is trained from annotation-supported outputs, Vanilla-DPO applies generic preference optimization under the matched backbone and inference protocol, and Balanced CPA-DPO applies pref-

Table 5: Distribution of annotated emotion–cause distances in ECPE-CN. Panel (a) reports the complete clause-distance distribution. Panel (b) aggregates the same annotations according to the evaluation grouping used throughout the paper. LOCAL contains pairs with $|e-k| \leq 1$, whereas LONG contains pairs with $|e-k| \geq 2$.

(a) Fine-grained distance distribution		
Distance	**# Gold Pairs**	**Proportion**
0	511	23.58%
1	1,342	61.93%
2	224	10.34%
≥ 3	90	4.15%
Total	**2,167**	**100.00%**
(b) Evaluation grouping		
Group	**# Gold Pairs**	**Proportion**
LOCAL: $\lvert e-k \rvert \leq 1$	1,853	85.51%
LONG: $\lvert e-k \rvert \geq 2$	314	14.49%

Table 6: Matched-protocol audit for the SFT, Vanilla-DPO, and Balanced CPA-DPO comparisons. "Matched" denotes exact agreement across the compared systems. The data-isolation check is performed fold-wise: preference construction uses only the corresponding training partition and does not access the held-out test partition.

Audit Item	**Result**
Qwen2.5-32B backbone configuration	**Matched (100%)**
Frozen emotion extractor	**Matched (100%)**
Fold-wise emotion predictions	**Matched (100%)**
Emotion-conditioned input template	**Matched (100%)**
Structured output format	**Matched (100%)**
Direct test-time evaluator	**Matched (100%)**
Test-time decoding configuration	**Matched (100%)**
Test examples exposed during preference construction	**0**
SFT vs. Vanilla-DPO emotion-prediction disagreement	**0**
Vanilla-DPO vs. Balanced CPA-DPO emotion-prediction disagreement	**0**

erence optimization using the counterfactually constructed chosen–rejected comparisons described in Sec. 3.

Table 6 summarizes the protocol audit.

Fixed upstream predictions. The zero-disagreement results in Table 6 directly verify that the matched end-to-end comparisons use identical upstream emotion predictions. Formally,

$$\Pr\left[\widehat{\mathcal{E}}_{\mathrm{SFT}}(D) \neq \widehat{\mathcal{E}}_{\mathrm{Vanilla}}(D)\right] = 0, \tag{18}$$

and

$$\Pr\left[\widehat{\mathcal{E}}_{\mathrm{Vanilla}}(D) \neq \widehat{\mathcal{E}}_{\mathrm{CPA}}(D)\right] = 0. \tag{19}$$

This control is important because the methodological intervention of Balanced CPA-DPO is restricted to emotion-conditioned cause selection. Accordingly, the matched performance differences cannot be explained by providing Balanced CPA-DPO with a stronger emotion extractor or a different set of predicted emotion clauses.

Fold-wise data isolation. The ECPE-CN evaluation follows the document-level ten-fold protocol described in the main paper. Within each fold, counterfactual generation, deterministic preprocessing, verifier scoring, eligibility filtering, hardest-valid selection, and preference formation operate

only on the corresponding training partition. No held-out test document or test annotation is supplied to the preference-construction pipeline. The measured test-exposure rate during preference construction is therefore zero.

Matched inference protocol. All matched systems are evaluated with the same original test documents, the same predicted target emotions, the same emotion-conditioned instruction template, the same structured JSON output schema, and the same direct decoding protocol.

For cause selection, the headline evaluator uses deterministic generation,

$$\texttt{do_sample=False}, \qquad \texttt{max_new_tokens} = 128. \tag{20}$$

The same structured-output parser is applied to SFT, Vanilla-DPO, and Balanced CPA-DPO.

No teacher model, frozen verifier, candidate generator, counterfactual-selection module, reranker, or gated decoding procedure is invoked to produce the reported headline results. The teacher and verifier are used only during offline preference-data construction.

Scope of the audit. These checks rule out several direct protocol confounds in the matched comparison, including differences in backbone capacity, upstream emotion predictions, inference input/output format, decoding configuration, and test-set exposure during preference construction. They do not eliminate ordinary optimization variation across folds. The latter is evaluated separately through the fold-level paired statistical analysis reported in Appendix D.12.

B Preference Construction and Implementation Invariants

The main paper presents the motivation and design of counterfactual preference construction. This appendix records the implementation-level details that determine the actual preference data: candidate generation, prompt interfaces, verifier scoring, eligibility, hardest-valid selection, position-balanced preference formation, and the structure of the chosen–rejected outputs.

Unless otherwise stated, the default verifier thresholds are

$$a_{\min} = 0.8, \qquad \tau = 0.3, \tag{21}$$

and the counterfactual generator requests

$$K = 3 \tag{22}$$

alternatives for each pair-level source record.

B.1 Candidate Generation and Deterministic Filtering

The preference builder operates on cause-targeted cause-selection records. Each record identifies a target emotion e and one annotation-supported target cause k_{gold}:

$$z = (D, e, k_{\text{gold}}, c_{\text{gold}}), \qquad c_{\text{gold}} = c_{k_{\text{gold}}}. \tag{23}$$

Let $\mathcal{G}(z)$ denote the ordered list of distinct annotation-supported cause indices associated with the same target emotion in the source annotation. For a single-cause record, $\mathcal{G}(z) = [k_{\text{gold}}]$.

Most final preference records can be matched to source tuples reconstructed under the single-cause source rule, while a small subset originates from target emotions with multiple annotated causes. For a multi-cause-derived record, k_{gold} denotes the particular annotation-supported cause targeted by generation and verifier screening, whereas $\mathcal{G}(z)$ may contain additional annotation-supported causes. Thus, counterfactual construction is cause-targeted, but the serialized chosen completion may retain the additional annotated causes after position remapping.

The generator uses GPT-4o with

$$T = 0.4 \tag{24}$$

and requests exactly three candidates. The implementation identifies the prompt as `counterfactual_generation_v2`. The API payload contains a single `user` message and does not add a separate `system` message. The implementation does not explicitly set `top_p`, `max_tokens`, a generation seed, or a JSON response-format option.

Teacher-output parsing. The returned content is first stripped of leading and trailing whitespace. If a Markdown fenced block is present, the parser splits the response on triple backticks, preferentially selects a block beginning with `json`, otherwise searches for a block beginning with "[", and then applies `json.loads`. The parsed top-level object must be a JSON list.

Each list element is converted to a string and then deterministically normalized. The builder:

1. removes empty strings;
2. strips leading numeric enumeration;
3. collapses repeated whitespace;
4. removes candidates whose normalized text exactly matches any normalized source clause;
5. removes duplicate candidates within the generated pool; and
6. retains at most the first three surviving candidates in generation order.

Because the annotation-supported cause is itself a source clause, the exact-source-clause filter also removes a verbatim copy of c_{gold}. We denote the surviving pool by

$$\mathcal{H}_{\text{rule}}(z) \subseteq \mathcal{H}(z). \tag{25}$$

If no candidate survives, the source record does not proceed to verifier-based selection.

Scope of deterministic filtering. The post-processing stage is structural rather than semantic. The implementation does not contain an independent deterministic candidate-length range, lexical-quality classifier, or semantic-relevance classifier. Participant preservation, contextual compatibility, and approximate length matching are instructions to the teacher rather than independently verified constraints.

B.2 Teacher Prompt: Executed Chinese Template and English Translation

The teacher prompt explicitly separates *causal core features*, which the model is instructed to disrupt, from *associative features*, which the model is instructed to preserve. This prompt-level decomposition guides candidate generation but should not be interpreted as an independently verified semantic constraint.

Executed Chinese prompt. The following Chinese template is the prompt actually sent to GPT-4o after sample-specific fields are filled:

```
你是情绪-原因分析与因果干预专家。
文档:
{doc_text}

已知:
- 情绪: [{emotion_idx + 1}] "{emo_sent}"
- 目标原因: [{cause_idx + 1}] "{cause_sent}"{context_note}

我们已拆解该原因的特征:
1. 因果核心特征（改了就不会触发情绪）:
{core_str}

2. 语义关联特征（保留它以维持语境）:
{assoc_str}

请生成 3 个\反事实原因候选"来替换原句:
要求:
1. 保留语义关联特征（人物、场景、时间不变）。
2. 移除/破坏因果核心特征（使其不再能触发该情绪）。
3. 风格长度与原句相近。

请只输出 JSON 数组:
["反事实句1", "反事实句2", "反事实句3"]
```

When the target emotion has other annotated causes, {context_note} is populated with the additional instruction

> 注意：该情绪可能还有其他原因句 [x], [y]，但请你【只针对】原因 [k] 进行修改，生成它的反事实版本。

and is empty otherwise. This detail is important for the final data provenance: multi-cause emotions can contribute pair-level preferences because the teacher is explicitly instructed to modify only the selected target cause while leaving other annotated causes outside the intervention target.

Faithful English translation. The following translation is included only for reader convenience; it is *not* a second prompt sent to the model.

```
You are an expert in emotion--cause analysis and causal
intervention.
Document:
{doc_text}

Known information:
- Emotion:  [{emotion_idx + 1}] "{emo_sent}"
- Target cause:  [{cause_idx + 1}] "{cause_sent}"
{context_note}

We have decomposed the target cause into the following
features:
1.  Causal core features (changing them should prevent the
emotion from being triggered):
{core_str}

2.  Semantically associated features (preserve them to
maintain contextual compatibility):
{assoc_str}

Generate three ``counterfactual cause candidates'' to
replace the original cause clause.
Requirements:
1.  Preserve the associated features, including
participants, scene, and time.
2.  Remove or disrupt the causal core features so that the
clause should no longer trigger the target emotion.
3.  Keep the style and length similar to the original
clause.

Return only a JSON array:
["counterfactual sentence 1", "counterfactual sentence 2",
"counterfactual sentence 3"]
```

For a multi-cause emotion, the additional note translates to:

```
Note:  this emotion may have other cause clauses [x]
and [y], but modify only cause [k] and generate the
counterfactual version of that cause.
```

Dynamic fields. The prompt fields have the following implementation meanings:

- doc_text: the original document, with each clause rendered as "[index] clause text";
- emotion_idx: the zero-based internal emotion index, exposed to the prompt after adding one;
- emo_sent: the target emotion-clause text;
- cause_idx: the zero-based internal target-cause index, exposed after adding one;

- `cause_sent`: the annotation-supported target-cause text;
- `core_str`: the extracted causal-core feature description;
- `assoc_str`: the extracted associative-feature description; and
- `context_note`: the optional multi-cause instruction above.

The generation instructions concerning preserved participants, scene, time, and approximate length are not re-verified by an independent semantic classifier after generation. Their role is to shape the candidate distribution supplied to the subsequent verifier-based screening stage.

B.3 Verifier Prompt and Target-Emotion Support Scoring

The frozen verifier g_ϕ does not classify a candidate clause in isolation. Instead, it evaluates whether the target emotion remains logically supported by the *complete document context*. This distinction is important for interpreting the construction score.

For compact notation, define the target-emotion support score

$$S(D, e) = P_{g_\phi}\left(\text{“是”} \mid D, e\right). \tag{26}$$

The implementation identifies the prompt as `emotion_support_yes_no_v2` and the scoring routine as `yes_no_token_probability_v2`.

Executed Chinese verifier prompt. For each document view, the following template is instantiated:

```
你是情绪逻辑分析专家。阅读以下文档:

{doc_text}

分析任务:
文档第 [{emotion_idx + 1}] 句的内容是: \{emotion_sent}"。
请判断: 结合上下文, 该句中表达的\{emotion_desc}"情绪是否逻辑合理且有
充分的原因支持?

请回答\是"或\否"。
答案:
```

Faithful English translation. The English translation below is explanatory only and is not used at runtime:

```
You are an expert in emotion-logic analysis.  Read the
following document:

{doc_text}

Task:
Clause [{emotion_idx + 1}] contains:  "{emotion_sent}".
Considering the full context, determine whether the
"{emotion_desc}" emotion expressed in this clause is
logically reasonable and sufficiently supported by causes
in the document.

Answer only ``Yes'' or ``No''.
Answer:
```

The runtime Chinese emotion descriptions are

$$\begin{aligned} \texttt{happiness} &\to \text{高兴}, \quad \texttt{sadness} \to \text{悲伤}, \\ \texttt{anger} &\to \text{愤怒}, \quad \texttt{fear} \to \text{恐惧}, \\ \texttt{disgust} &\to \text{厌恶}, \quad \texttt{surprise} \to \text{惊讶}. \end{aligned} \tag{27}$$

If an emotion label is not present in this mapping, the original label text is used.

Binary token-probability scoring. The verifier does not autoregressively generate the answer. It reads the final-position logits for the single tokens "是" and "否". The tokenizer is required to map each answer string to exactly one token; otherwise the scoring implementation terminates with a tokenizer error.

Let these logits be $\ell_{是}$ and $\ell_{否}$. The support score is

$$S(D, e) = \frac{\exp(\ell_{是})}{\exp(\ell_{是}) + \exp(\ell_{否})}. \tag{28}$$

Original and counterfactual document views. For the original document,

$$A_{\text{orig}} = S(D, e). \tag{29}$$

For generated candidate $c_{\text{trap}}^{(j)}$, the builder copies the original clause sequence and replaces the target cause at its original position:

$$D_{\text{cf}}^{(j)} = D_{c_{\text{gold}} \to c_{\text{trap}}^{(j)}}. \tag{30}$$

The same verifier prompt is then rebuilt from the complete replaced document, yielding

$$A_{\text{cf}}^{(j)} = S(D_{\text{cf}}^{(j)}, e). \tag{31}$$

The support reduction induced by the replacement is

$$\Delta_j = A_{\text{orig}} - A_{\text{cf}}^{(j)}. \tag{32}$$

Thus, Δ_j measures the drop in learned target-emotion support after replacing the annotation-supported cause with candidate j. It is an intervention-inspired verifier contrast, not a formal causal effect.

Verifier inference configuration. Verifier scoring is deterministic. Inputs are left-padded and truncated to a maximum sequence length of 2,048 tokens. The verifier is loaded using 4-bit NF4 quantization with double quantization and bfloat16 computation.

The verifier checkpoint is configured through `DECC_SFT_MODEL_PATH`; the repository default resolves to `/data/final_merged`. This path is an implementation location rather than a stable public checkpoint identifier. Accordingly, we denote by g_ϕ the frozen checkpoint actually used by the construction pipeline and do not infer a more specific model identity from the path name alone. The implementation does not persist or invoke an additional reference scorer s_ψ.

B.4 Eligibility Correctness

A generated candidate is eligible only if the original document receives sufficient target-emotion support and replacing the annotation-supported cause induces a sufficiently large support reduction:

$$A_{\text{orig}} \geq 0.8, \qquad \Delta_j \geq 0.3, \tag{33}$$

with all required scores present and finite.

The verifier-eligible set is

$$\mathcal{V}(z) = \left\{ c_{\text{trap}}^{(j)} \in \mathcal{H}_{\text{rule}}(z) \;\middle|\; A_{\text{orig}} \geq 0.8,\ \Delta_j \geq 0.3 \right\}. \tag{34}$$

We use *valid* operationally: it means that the candidate satisfies the implemented construction criteria, not that the candidate has been assigned an objective causal status.

If

$$\mathcal{V}(z) = \varnothing, \tag{35}$$

the source record produces no preference example.

Interpretation. The eligibility stage first requires the annotation-supported relation to be strongly supported under the original document. It then requires the replacement to reduce that support by at least 0.3. The reliability constraint is therefore applied before any candidate-hardness ranking.

B.5 HARDEST-VALID SELECTION CORRECTNESS

Among eligible candidates, the implementation retains the replacement that leaves the *highest residual target-emotion support*:

$$c^{\star}_{\text{trap}} = \arg \max_{c^{(j)}_{\text{trap}} \in \mathcal{V}(z)} A^{(j)}_{\text{cf}}. \tag{36}$$

Because A_{orig} is fixed across candidates generated for the same source record,

$$\arg \max_{c_j \in \mathcal{V}(z)} A^{(j)}_{\text{cf}} = \arg \min_{c_j \in \mathcal{V}(z)} \Delta_j. \tag{37}$$

Thus, "hardest-valid" is verifier-relative: among replacements that already satisfy the eligibility constraints, it selects the one whose replaced document still receives the greatest residual support for the target emotion. It does *not* mean the replacement with the largest support drop, an objectively hardest negative, or the candidate closest to a policy decision boundary.

If candidates are tied under the implemented score ordering, the smaller candidate index is selected deterministically.

The construction therefore follows

$$\boxed{\text{eligibility first } \rightarrow \text{ hardness second}}. \tag{38}$$

B.6 POSITION-BALANCED PREFERENCE FORMATION

Replacement is used for verifier screening and for one of the two position-balanced preference orientations. After hardest-valid selection, the builder forms the actual DPO input with the selected counterfactual, the targeted gold cause, and one auxiliary dummy clause.

Let

$$D = (c_1, \ldots, c_n). \tag{39}$$

The two appended clauses occupy exactly positions $n+1$ and $n+2$. Let

$$b \sim \text{Bernoulli}(0.5) \tag{40}$$

select the semantic orientation. Independently, let

$$\rho \sim \text{Bernoulli}(0.5) \tag{41}$$

select the exact order of the two appended clauses. With the default

$$\texttt{swap_prob} = 0.5, \tag{42}$$

let $\text{Perm}_\rho(u, v)$ denote either (u, v) or (v, u) according to ρ. The appended clauses are

$$(a_{n+1}, a_{n+2}) = \begin{cases} \text{Perm}_\rho(c^{\star}_{\text{trap}}, c_{\text{dummy}}), & b = 0, \\ \text{Perm}_\rho(c_{\text{gold}}, c_{\text{dummy}}), & b = 1. \end{cases} \tag{43}$$

The orientation-specific document is

$$\widetilde{D}_b = \begin{cases} \text{Append}(D, a_{n+1}, a_{n+2}), & b = 0, \\ \text{Append}(D_{c_{\text{gold}} \rightarrow c^{\star}_{\text{trap}}}, a_{n+1}, a_{n+2}), & b = 1. \end{cases} \tag{44}$$

For $b = 0$, the targeted gold cause remains at its original position and the trap is appended. For $b = 1$, the trap occupies the original targeted gold-cause position and that gold cause is appended. Let k^{+}_{gold} and k^{+}_{trap} denote the resulting positions of these two semantic roles. Exactly one is original and the other lies in the two appended slots:

$$k^{+}_{\text{gold}}, k^{+}_{\text{trap}} \in \{k_{\text{gold}}, n+1, n+2\}, \qquad \{k^{+}_{\text{gold}}, k^{+}_{\text{trap}}\} = \{k_{\text{gold}}, k_{\text{app}}\}, \tag{45}$$

where $k_{\text{app}} \in \{n+1, n+2\}$. The emotion index remains unchanged, $e^{+} = e$.

Dummy-clause construction. The dummy clause is not generated by GPT-4o or any other model. The construction script contains a fixed pool $\mathcal{D}_{\text{dummy}}$ of five hard-coded Chinese filler templates. Examples include （无关信息） and （背景补充信息，与因果无直接关系）.

For each preference record, `choose_dummy()` draws from this pool using `random.choice`. The script uses random seed 42 by default, so the sequence of dummy choices is reproducible given the same record order.

The selected template is compared with original document clauses after removing leading clause numbering. If it exactly matches a source clause, the builder resamples, for at most 20 attempts. If all attempts collide, the implementation falls back to （无关信息） without another collision check. The procedure therefore attempts to avoid exact duplication with an original clause but does not provide an absolute non-duplication guarantee.

Because only five fixed templates are available, dummy clauses are deliberately reused across many preference records. The dummy is neither verifier-ranked nor used as the rejected target. Its role is structural: together with the semantic clause selected for the current orientation, it creates two appended positions while the gold/trap roles are balanced across original and appended positions.

Index convention. The builder stores original positions internally with zero-based indices, whereas indices exposed to the prompt and structured prediction interface are one-based. Appending clauses does not shift the indices of original clauses. For $b = 1$, however, the targeted gold cause is relocated to one of the appended positions; its index is therefore remapped in the chosen completion, while any other annotation-supported causes retain their original positions.

Preference outputs. The shared augmented input is

$$x^+ = x(\widetilde{D}_b, e). \tag{46}$$

Let $\mathcal{G}_b^+(z)$ denote $\mathcal{G}(z)$ after remapping the targeted gold index when $b = 1$; all other annotated cause indices remain unchanged. The preferred completion is

$$y_w = \{\texttt{emotion_idx} : e,\ \texttt{cause_idxs} : \mathcal{G}_b^+(z)\}, \tag{47}$$

whereas the rejected completion contains the selected counterfactual index:

$$y_l = \{\texttt{emotion_idx} : e,\ \texttt{cause_idxs} : [k^+_{\text{trap}}]\}. \tag{48}$$

For the dominant single-cause case, $\mathcal{G}_b^+(z) = [k^+_{\text{gold}}]$. For multi-cause-derived records, the chosen completion may retain additional annotation-supported causes in $\mathcal{G}_b^+(z)$, while generation and verifier screening remain targeted to k_{gold}. The construction is therefore cause-targeted rather than universally one-cause-versus-one-trap, and it does not explicitly compare alternative complete cause sets.

Scope of position balancing. Balancing the semantic roles across original and appended positions removes the deterministic mapping from appended status to preference label. The measured position-only and full-metadata probes in Section 4.4 evaluate this construction control. It does not remove positional information from the input or establish position invariance.

B.7 Controlled Preference Integrity

Every chosen–rejected comparison is formed under the same augmented document:

$$(x^+, y_w, y_l). \tag{49}$$

The builder enforces the structural properties in Table 7.

Thus, all chosen and rejected outputs share the same augmented context, target emotion, and structured output schema. Equal answer cardinality is not a universal invariant: 5,994 of the 6,178 final preferences (97.02%) have equal chosen and rejected cardinality. The remaining 184 records are multi-cause-derived cases in which the chosen completion retains multiple distinct annotation-supported causes while the rejected completion contains the selected trap.

Table 7: Structural controls and observed properties of CPA preference construction.

Integrity Check	**Status**
Shared preference input, $x_w = x_l = x^+$	Enforced
Same target emotion	Enforced
Same structured JSON schema	Enforced
Semantic gold/trap roles balanced across original and appended positions	Enforced
Exactly two clauses appended	Enforced
Final appended semantic clause and dummy occupy $n+1$ or $n+2$	Enforced
Same answer cardinality	Observed: 97.02%

Training scope. The final preference dataset contains 6,178 cause-targeted preference records. Of these, 5,936 can be matched to the reconstructed single-cause source pool, while 242 (3.92%) originate from target emotions with multiple annotated causes. Among the 242 multi-cause-derived records, 184 retain multiple distinct annotation-supported causes in the chosen completion, whereas 58 deduplicate to one distinct cause. The data are therefore dominated by single-cause-derived records but are not exclusively single-cause in their serialized chosen outputs.

Generation and verifier screening remain targeted to one annotation-supported cause at a time. The chosen completion can preserve additional annotated causes, but the current objective does not explicitly compare alternative complete cause sets. Complete set-valued cause recovery is therefore only partially represented by the current supervision. The corresponding downstream slice analysis is reported in Appendix D.7.

B.8 Summary of Construction Invariants

The implementation-level construction can be summarized as

$$\boxed{\begin{array}{c}\text{GPT-4o generation} \rightarrow \text{deterministic cleanup} \\ \rightarrow \text{document-level verifier eligibility} \\ \rightarrow \text{hardest-valid selection} \\ \rightarrow \text{position-balanced preference formation}\end{array}}. \tag{50}$$

The key invariants are:

1. GPT-4o is prompted for $K = 3$ alternatives at temperature 0.4;
2. deterministic post-processing removes malformed, empty, duplicate, and exact-source-clause outputs but does not independently certify the teacher's semantic instructions;
3. verifier scoring measures target-emotion support under the complete original or replacement document rather than candidate-cause probability in isolation;
4. eligibility requires $A_{\text{orig}} \geq 0.8$ and $\Delta_j \geq 0.3$;
5. hardest-valid selection maximizes residual $A_{\text{cf}}^{(j)}$ within the eligible set;
6. semantic gold/trap roles are balanced across the original targeted cause position and the appended positions, while the exact order of the appended semantic clause and dummy is randomized independently at probability 0.5; and
7. generation and verifier screening remain cause-targeted, while the serialized chosen completion may retain additional annotation-supported causes for multi-cause-derived records; equal answer cardinality is therefore not a universal invariant, and complete set-valued preference optimization is not implemented.

These are implementation and supervision properties rather than independent downstream performance claims. Their behavioral consequences are evaluated in the main-paper ablations and shortcut-sensitive experiments.

Table 8: Confirmed training configuration of the frozen emotion extractor.

Setting	Value
Training epochs	2
Learning rate	2×10^{-4}
Per-device batch size	2
Gradient accumulation	4
Warmup ratio	0.03
Maximum sequence length	2048
Fold seed	$42 + \text{fold}$
LoRA rank	64
LoRA scaling factor	16
LoRA dropout	0.05
LoRA targets	attention + feed-forward projections
Quantization	4-bit NF4
Checkpoint selection	fixed two-epoch schedule

C Confirmed Training Configuration and Inference

This appendix reports only training and inference settings that can currently be tied to the implemented matched pipeline without inferring unresolved hyperparameters from neighboring stages. In particular, the frozen emotion-extractor configuration and the DPO-stage configuration are documented below. The DPO policy is initialized from the corresponding ECE SFT checkpoint, but the complete ECE-SFT trainer configuration is intentionally not reconstructed here from incomplete provenance.

C.1 Frozen Emotion Extractor

All matched SFT, Vanilla-DPO, and Balanced CPA-DPO systems reuse the same fold-specific emotion extractor and the same resulting emotion predictions.

For each fold, the extractor is initialized from the corresponding Qwen2.5-32B ECPE-CN SFT checkpoint and adapted with a fold-specific emotion LoRA module. It is trained only on the corresponding training partition for two epochs with the configuration in Table 8.

The extractor does not use development-based checkpoint selection. At inference time it generates one-based emotion-clause indices in structured JSON using deterministic decoding:

$$\texttt{do_sample=False}, \qquad T = 0, \qquad \texttt{max_new_tokens} = 256. \tag{51}$$

Predicted indices are restricted to the valid clause range. These fold-wise predictions are then frozen and reused unchanged by the matched cause-selection systems.

C.2 ECE Preference Optimization

The main matched cause-selection experiments use the same Qwen2.5-32B backbone configuration. The DPO policy is initialized from the corresponding ECE SFT cause-selector checkpoint, and that SFT policy is also used as the frozen reference:

$$\pi_{\text{ref}} = \pi_{\text{SFT}}. \tag{52}$$

Only the ECE adapter is updated during preference optimization. The upstream emotion extractor remains fixed.

The confirmed DPO-stage configuration is shown in Table 9.

The methodological novelty does not lie in modifying the DPO objective. It lies in the construction of the chosen–rejected supervision described in Appendix B.

C.3 Final Checkpoint Selection

All headline and supplementary CPA-DPO results in this appendix use the raw post-DPO policy produced by the final position-balanced training run. No SFT–DPO adapter interpolation, checkpoint

Table 9: Confirmed DPO-stage configuration for the emotion-conditioned cause selector.

Setting	Value
Learning rate	3×10^{-6}
DPO coefficient β	0.1
Per-device batch size	1
Gradient accumulation	24
Training epochs	1
Maximum sequence length	512
Warmup ratio	0.1
Fold seed	$42 + \text{fold}$
Quantization	4-bit NF4
Computation	bfloat16
Optimizer family	paged AdamW
Gradient checkpointing	Enabled

Table 10: Normalized computational characteristics relative to Vanilla-DPO.

Quantity	CPA / Vanilla-DPO
Trainable-parameter ratio	1.00×
Peak-VRAM ratio	1.02×
DPO-stage wall-clock ratio	1.04×
Inference-latency ratio	1.008×
Extra test-time module	None

averaging, reranking, or auxiliary post-training mixture is used to obtain the reported final values. The SFT checkpoint initializes the DPO policy and is also used as the frozen reference policy during optimization, but the deployed CPA-DPO checkpoint is the direct post-DPO policy.

C.4 Computational Characteristics

Balanced CPA-DPO adds offline computation for GPT-4o generation, verifier scoring, eligibility filtering, hardest-valid selection, and preference formation. These operations are absent from the deployed inference pipeline.

Table 10 reports the normalized measurements available for the matched Balanced CPA-DPO and Vanilla-DPO systems. The DPO-stage wall-clock ratio excludes offline teacher generation and verifier scoring.

The latency difference is a measured system-level quantity rather than an algorithmic complexity claim. The method changes training-time supervision while preserving a single direct-generation ECE model at deployment.

C.5 Direct Inference Protocol

All headline SFT, Vanilla-DPO, and Balanced CPA-DPO results use the same direct structured-generation evaluator.

Given an original test document D, the frozen emotion extractor first produces

$$\widehat{\mathcal{E}}(D). \tag{53}$$

For each predicted emotion $\hat{e}$, the cause selector receives

$$x(D, \hat{e}) \tag{54}$$

constructed from the original test document, with no training-time counterfactual or dummy clause.

Cause-selection generation is deterministic:

$$\texttt{do_sample=False}, \qquad \texttt{max_new_tokens} = 128. \tag{55}$$

The evaluator scans the generated text and returns the last valid JSON object containing the required structured fields. The same evaluator and parser are used for matched SFT, Vanilla-DPO, and Balanced CPA-DPO comparisons.

The repository also contains a gated evaluation utility, but that utility is not used for the reported headline results. No teacher generation, verifier scoring, counterfactual augmentation, reranking, confidence gate, or generate-and-rerank branch is invoked at test time.

C.6 Structured-Output Robustness

Across the evaluated direct-generation outputs,

$$99.8\% \tag{56}$$

are parseable as valid JSON under the evaluation protocol, and

$$99.7\% \tag{57}$$

contain valid clause indices.

These quantities measure output well-formedness rather than ECPE accuracy: a syntactically valid JSON prediction can still select an incorrect cause. The qualitative error audit in Appendix E.1 separately reports the fraction of sampled errors whose primary failure mode is formatting or indexing.

C.7 Reproducibility Scope

The implementation separates three roles:

1. GPT-4o is used offline to propose counterfactual candidates;
2. the frozen document-level emotion-support verifier g_ϕ is used offline to screen and rank candidates; and
3. the final ECE model performs direct structured generation at test time.

The verifier checkpoint is loaded through the configuration described in Appendix B.3. Because the repository-local default path is not itself a stable public model identifier, we do not infer an exact checkpoint identity beyond the artifact actually used by the construction pipeline.

No additional persisted reference scorer is required by preference construction. The DPO reference policy is instead the frozen ECE SFT policy, $\pi_{\text{ref}} = \pi_{\text{SFT}}$.

D Additional Experimental Evidence

This appendix provides supplementary quantitative analyses supporting the empirical findings reported in the main paper. The primary effectiveness results, shortcut-sensitive evidence, cause/non-cause discrimination results, construction-validity evidence, and core component ablations remain in the main text. Here, we report complete backing values, threshold sensitivity, preference-data composition and structural audits, human evaluation, objective control, single- versus multi-cause slices, backbone generalization, cross-domain transfer, extended hard-negative diagnostics, and fold-level statistical tests.

Throughout this appendix, we use the same long-distance definition as in the main paper:

$$\text{LONG} : |e - k| \geq 2. \tag{58}$$

The analyses support the narrower claim that CPA-DPO reduces reliance on positional shortcuts. They do not imply that positional information is eliminated or that the resulting model is position invariant.

D.1 Complete Shortcut-Sensitive Results

The main paper presents the central shortcut-sensitive findings. Table 11 reports the final matched SFT, Vanilla-DPO, and Balanced CPA-DPO values.

Table 11: Complete shortcut-sensitive evaluation on ECPE-CN. LONG contains emotion–cause pairs with $|e - k| \geq 2$. Worst Drop is the worst relocation degradation relative to the unperturbed input; lower is better.

Method	Overall F1	LONG F1	Rebalanced	Worst Drop ↓
SFT	73.07	46.39	61.80	6.16
Vanilla-DPO	83.62	70.42	74.65	4.27
Balanced CPA-DPO	**85.43**	**75.46**	**79.95**	**2.72**

Table 12: Sensitivity to verifier eligibility thresholds for the final position-balanced CPA-DPO construction. The default configuration $(A_{\min}, \tau) = (0.8, 0.3)$ is shown in bold. HN denotes Constructed HN Hit@1 rather than Independent HN Accuracy.

$A_{\min}$	τ	Overall	LONG	Rebalanced	HN Hit@1
0.7	0.2	84.78	73.12	77.94	92.35
0.7	0.3	85.15	74.35	78.91	92.93
0.7	0.4	84.91	73.45	78.09	92.68
0.8	0.2	85.24	74.61	79.18	93.08
0.8	**0.3**	**85.43**	**75.46**	**79.95**	**93.45**
0.8	0.4	85.18	74.55	79.12	93.22
0.9	0.2	84.87	73.36	78.04	92.75
0.9	0.3	85.08	74.22	78.83	93.15
0.9	0.4	84.75	73.45	78.21	92.82

Relative to Vanilla-DPO, Balanced CPA-DPO improves aggregate ECPE-CN F1 by

$$85.43 - 83.62 = +\mathbf{1.81}, \tag{59}$$

whereas the improvement on LONG examples increases to

$$75.46 - 70.42 = +\mathbf{5.04}. \tag{60}$$

The same amplification is observed under position-rebalanced evaluation: Balanced CPA-DPO improves Vanilla-DPO from 74.65 to 79.95 F1, corresponding to

$$79.95 - 74.65 = +\mathbf{5.30}. \tag{61}$$

Relative to the Overall gain, the corresponding point-estimate amplification is $+3.23$ F1 for LONG and $+3.49$ F1 for Rebalanced-CN.

Controlled position perturbations provide complementary evidence. The worst relocation degradation decreases from 4.27 points for Vanilla-DPO to 2.72 points for Balanced CPA-DPO. The remaining non-zero degradation is important: Balanced CPA-DPO remains sensitive to positional changes. Together with the larger LONG and Rebalanced-CN gains, these results are consistent with reduced reliance on the benchmark's dominant positional regularity rather than shortcut elimination or position invariance.

D.2 Threshold Sensitivity

The final threshold sensitivity analysis evaluates the verifier eligibility settings around the default configuration. The default verifier thresholds are

$$A_{\min} = 0.8, \qquad \tau = 0.3. \tag{62}$$

To test whether the observed behavior depends on a narrowly tuned configuration, we evaluate

$$A_{\min} \in \{0.7, 0.8, 0.9\}, \qquad \tau \in \{0.2, 0.3, 0.4\}. \tag{63}$$

Table 12 reports aggregate ECPE-CN F1, LONG F1, Rebalanced-CN F1, and Constructed HN Hit@1.

Within the evaluated grid, the default configuration achieves the strongest result on all four reported measures, while neighboring configurations remain competitive. Changing either threshold by 0.1

Table 13: Composition of the final CPA preference dataset. "Single-source matched" denotes records matching the reconstructed 18,056-tuple source pool under the single-cause source-key rule.

Preference-Data Group	Count	Share
Single-source matched	5,936	96.08%
Multi-cause-derived	242	3.92%
Final preference data	**6,178**	**100.00%**

Table 14: Construction-label leakage audit for the final Balanced CPA data.

Diagnostic	Result
$P(\text{appended} \mid \text{chosen})$	50.2%
$P(\text{appended} \mid \text{rejected})$	49.8%
Position-only probe accuracy	50.6%
Full construction-metadata probe accuracy	52.1%
Chance baseline	50.0%

does not produce an abrupt performance collapse. The evaluated region therefore exhibits a broad intermediate optimum rather than a sharp threshold cliff. We do not interpret $(0.8, 0.3)$ as universally optimal outside the evaluated setting.

D.3 Final Preference-Data Composition and Structural Audit

We next characterize the chosen–rejected preference data used for the reported CPA-DPO experiments. These statistics describe training-data composition and construction properties rather than downstream model performance.

The final preference dataset contains

$$N_{\text{pref}} = 6{,}178 \tag{64}$$

records across the ten ECPE-CN folds. An offline source-key audit reconstructs

$$N_{\text{single-source}} = 18{,}056 \tag{65}$$

source tuples under the code's single-cause reconstruction rule.

Of the final 6,178 records, 5,936 match this reconstructed source pool, whereas 242 are derived from annotations containing multiple cause entries.

The 242 multi-cause-derived records require a more precise description of the construction. Among them, 184 contain multiple distinct annotated cause indices and retain the complete annotated cause set in the chosen completion. The remaining 58 contain multiple annotation entries that deduplicate to a single distinct cause.

Consequently, the final preference data should not be described as universally one-cause-versus-one-trap comparisons. The construction procedure is *cause-targeted*: generation and screening operate with respect to one targeted annotation-supported cause, while the serialized chosen completion can retain multiple annotated causes.

Across all 6,178 records, 5,994 preferences contain one chosen cause and 5,994 have equal chosen and rejected answer cardinality, corresponding to

$$\frac{5{,}994}{6{,}178} = 97.02\%. \tag{66}$$

Position-balanced construction controls. The final-record audit measures whether semantic role is associated with appended status or other construction metadata. The confirmed diagnostics are shown below.

The final-record audit further verifies the structural controls.

No dummy/original-clause collision or dummy/trap collision is observed in the final records. These checks support the controlled-construction claim without assuming universal one-cause output structure. They do not establish that all possible construction artifacts are absent.

Table 15: Structural audit of the final 6,178 CPA preference records.

Property	Verified Share
Same output schema	100.00%
Same target emotion	100.00%
Chosen cause index valid	100.00%
Rejected cause index equals selected trap role	100.00%
Gold/trap roles balanced across original and appended positions	100.00%
Exactly two clauses appended	100.00%
Trap metadata matches augmented prompt	100.00%
One chosen cause	97.02%
Same answer cardinality	97.02%

Table 16: Distance distribution of the reconstructed 18,056-tuple single-cause preference source pool. LONG is defined as $|e-k| \geq 2$.

Distance Group	Source Tuples	Share
$\lvert e-k\rvert = 0$	4,455	24.67%
$\lvert e-k\rvert = 1$	11,351	62.87%
$\lvert e-k\rvert = 2$	1,593	8.82%
$\lvert e-k\rvert \geq 3$	657	3.64%
LOCAL: $\lvert e-k\rvert \leq 1$	15,806	87.54%
LONG: $\lvert e-k\rvert \geq 2$	2,250	12.46%
Total	**18,056**	**100.00%**

The 18,056 reconstructed source tuples and the 6,178 final preference records are not directly interchangeable populations. The former is defined by the single-cause source reconstruction rule, whereas the latter additionally contains multi-cause-derived records. Accordingly, we do not report $6{,}178/18{,}056$ as a global preference retention rate.

The current retained-row artifacts preserve row-level A_{orig} and Δ fields, but do not preserve the complete candidate-level scoring history or the full eligible set $\mathcal{V}(z)$ for every final record. We therefore do not report final-set distributions of $|\mathcal{V}(z)|$, candidate-level rejection rates, selected $A^{\star}_{\text{cf}}$ distributions, or tie statistics.

D.4 Preference-Data Diagnostics and Human Audit

We complement the composition and structural audit with two diagnostics: locality composition and a manual audit of retained preferences. These analyses characterize the supervision used by CPA-DPO rather than replacing downstream ECPE evaluation.

D.4.1 Locality Composition

Appendix A.1 reports locality statistics over the original ECPE-CN annotated pairs. The preference-construction source audit uses a different statistical object: 18,056 fold-expanded source tuples reconstructed under the code's single-cause source rule.

Their distance distribution is shown in Table 16.

The reconstructed source pool is therefore strongly locality-skewed: 87.54% of source tuples satisfy $|e-k| \leq 1$, whereas 12.46% satisfy the LONG criterion.

The final preference set remains strongly locality-skewed as well. Table 17 compares composition shares between the reconstructed source pool and the final preference data.

Within the 5,936 final records matching the reconstructed single-cause source pool, 5,672 are LOCAL and 264 are LONG. Within the 242 multi-cause-derived records, 191 are LOCAL and 51 are LONG.

Table 17: Locality composition of the reconstructed source pool and final CPA preference data. Percentages are composition shares within each population and should not be interpreted as retention rates.

Population	LOCAL	Share	LONG	Share
Reconstructed source pool	15,806	87.54%	2,250	12.46%
Final preference data	5,863	94.90%	315	5.10%

Table 18: Human audit of 180 retained CPA preference records.

Audit Criterion	Result
Contextually plausible	91.1%
Gold cause clearly preferable	93%
Trap judged as a plausible cause	6%
Inter-annotator Cohen's κ	0.78

These values show that the final supervision remains dominated by local emotion–cause relations. The shortcut-sensitive improvements in the main paper therefore do not arise from simply constructing a training set dominated by long-distance examples. Because the source and final populations are not identical, we treat these numbers only as composition statistics and do not infer group-specific retention rates from them.

D.4.2 Human Audit of Retained Preferences

We additionally conduct a manual audit of $N = 180$ retained preference records. The audit examines whether the generated counterfactual remains plausible in context, whether the annotation-supported cause is clearly preferable, and whether the generated trap could itself reasonably still be judged as a cause of the target emotion.

Most generated alternatives remain contextually plausible (91.1%), supporting their use as non-trivial rejected alternatives rather than arbitrary negatives. At the same time, the annotation-supported cause is judged clearly preferable in 93% of audited comparisons, while only 6% of traps are judged as plausible causes themselves. The inter-annotator agreement is

$$\kappa = 0.78. \tag{67}$$

These results support the practical reliability of the retained preference supervision. They are used as construction diagnostics rather than as a replacement for downstream ECPE evaluation.

The final verifier diagnostics report eligibility coverage of 34% for LOCAL records and 29% for LONG records, and 35% for single-cause records versus 25% for multi-cause records. The lower coverage in the latter two slices is consistent with a conservative document-level support criterion rather than a slice collapse.

An additional audit of filtered-out candidates provides complementary evidence for the filter: the gold cause is judged clearly preferable in 74% of cases, and the trap is judged a plausible cause in 24% of cases, compared with 93% and 6%, respectively, among retained comparisons. The retained audit uses $N = 180$ records as reported above; the filtered-out summary is used only at the confirmed aggregate-percentage level. These figures characterize the verifier operating boundary and do not treat the verifier as an oracle.

D.5 Objective Control

The balanced preference data can also be optimized with a pairwise ranking objective. On the same evaluation slices, the ranking objective reaches 84.45 Overall F1, 73.82 LONG F1, and 90.31 Independent HN Accuracy, whereas standard DPO reaches 85.43, 75.46, and 91.35, respectively. The corresponding DPO advantages are +0.98 Overall, +1.64 LONG, and +1.04 Independent-HN points. This comparison supports positioning CPA-DPO primarily as a supervision-construction method rather than as a new preference objective; it does not show that DPO is necessary.

Table 19: Performance on single-cause and multi-cause test slices. Exact-set Accuracy for the multi-cause setting requires the complete predicted cause set to match the annotation.

Method	Single-Cause F1	Multi-Cause F1	Multi-Cause Exact-Set Acc.
SFT	74.80	62.10	61.20%
Vanilla-DPO	85.10	74.30	73.50%
Balanced CPA-DPO	**87.15**	**75.02**	**74.45%**

Table 20: Cross-backbone robustness/generalization on ECPE-CN for final position-balanced CPA-DPO. Δ_{Overall} and Δ_{LONG} denote gains over the corresponding Vanilla-DPO baseline.

Backbone	SFT	Vanilla-DPO	CPA-DPO	Δ_{Overall}	Δ_{LONG}
Qwen2.5-7B	68.74	78.26	**79.91**	+1.65	+4.21
Qwen2.5-14B	70.86	80.62	**82.31**	+1.69	+4.57
Llama-3.x-8B	67.83	77.14	**78.66**	+1.52	+3.88
Qwen2.5-32B	73.07	83.62	**85.43**	+1.81	+5.04

D.6 Final Checkpoint Convention

All final CPA-DPO numbers reported in the main paper and this appendix are measured from the raw post-DPO checkpoint. Adapter interpolation is not part of the final method, checkpoint selection, or inference path. This convention is kept explicit here because earlier experiment bookkeeping used an interpolation label for some runs that are reported here under their confirmed final position-balanced, raw post-DPO configuration.

D.7 Single-Cause and Multi-Cause Slices

The preference construction is cause-targeted rather than a complete set-valued preference objective. Generation and verifier screening operate with respect to one targeted annotation-supported cause. For most final records, this also yields a one-cause chosen completion. However, the final data contain 242 multi-cause-derived records: 184 retain multiple distinct annotated causes in the chosen completion, whereas 58 deduplicate to one distinct cause.

Thus, the final preference data are not uniformly one-cause-versus-one-trap comparisons. At the same time, the current construction does not explicitly compare alternative complete cause sets. We therefore evaluate downstream single-cause and multi-cause examples separately.

Relative to Vanilla-DPO, Balanced CPA-DPO improves the single-cause slice by $+2.05$ F1 points. The gain on multi-cause instances is smaller at $+0.72$ F1, while multi-cause exact-set Accuracy improves by $+0.95$ percentage points.

The positive but smaller multi-cause gain is consistent with the scope of the current supervision. Multi-cause-derived chosen completions preserve the available annotated cause set, but preference construction still targets one cause when generating and screening the counterfactual alternative. The method therefore provides only partial supervision for complete set-valued cause recovery.

We consequently treat complete multi-cause set modeling as a remaining extension rather than claiming that the current preference construction fully solves this setting.

D.8 Cross-Backbone Robustness and Generalization

We evaluate whether the benefit of the final position-balanced construction persists across backbone scales and model families. All CPA-DPO entries in Table 20 use the final position-balanced construction, the raw post-DPO policy, and no adapter interpolation.

Across all four evaluated backbones, CPA-DPO improves Vanilla-DPO by 1.52–1.81 Overall F1. The corresponding LONG gains are consistently larger, ranging from 3.88 to 5.04 points. The same shortcut-sensitive pattern therefore persists across the evaluated Qwen scales and the Llama

Table 21: Cross-domain results on RECCON-DD and RECCON-IEMOCAP. $F1_{\text{pos}}$ and $F1_{\text{neg}}$ denote class-wise F1, and Macro-F1 is their macro-averaged summary. All CPA-DPO entries are from the final position-balanced run.

Dataset	Method	$\mathbf{F1}_{\text{pos}}$	$\mathbf{F1}_{\text{neg}}$	Macro-F1
RECCON-DD	SFT	86.80	92.60	89.70
	Vanilla-DPO	88.30	94.40	91.35
	Balanced CPA-DPO	**89.60**	**96.20**	**92.90**
RECCON-IEMOCAP	SFT	72.80	93.50	83.15
	Vanilla-DPO	75.20	94.70	84.95
	Balanced CPA-DPO	**77.00**	**96.20**	**86.60**

Table 22: Extended hard-negative diagnostics for the final position-balanced CPA-DPO system. Constructed Hit@1 and Independent HN Accuracy are distinct metrics: the former evaluates construction-coupled comparisons, whereas the latter uses separately constructed hard negatives.

Method	Constructed Hit@1	Constructed Margin	Independent HN Acc.	Independent Margin	CF Cons.	Intervention Acc.
SFT	84.37	5.10	79.85	4.18	82.14	80.75
Random-DPO	89.12	7.65	84.10	5.96	86.73	85.44
Vanilla-DPO	91.30	10.20	87.92	8.41	90.88	89.95
Balanced CPA-DPO	**93.45**	**11.58**	**91.35**	**10.26**	**94.31**	**93.64**

backbone. These results provide robustness/generalization evidence across the tested model scales and families; they do not establish backbone-independent behavior beyond the evaluated models.

D.9 Cross-Domain Transfer on RECCON

We further document whether the ordering observed in the final position-balanced run persists in a different conversational emotion-cause setting. Table 21 reports results on RECCON-DD and RECCON-IEMOCAP using class-wise positive and negative F1 together with Macro-F1.

The final CPA-DPO ordering

$$\text{SFT} < \text{Vanilla-DPO} < \text{CPA-DPO} \tag{68}$$

is preserved on both datasets. Relative to Vanilla-DPO, the Balanced CPA-DPO improves Macro-F1 by $+1.55$ points on RECCON-DD and by $+1.65$ points on RECCON-IEMOCAP.

These results provide supporting evidence that the benefit of task-specific counterfactual preference construction is not confined to the primary ECPE-CN benchmark. We deliberately do not use RECCON as direct evidence for the positional shortcut claim, because the corresponding locality distribution and stress-test construction differ from those used for ECPE-CN. The cross-domain experiment therefore supports transfer of the overall method ordering rather than a specific claim about positional shortcut reduction.

D.10 Extended Hard-Negative Diagnostics

The main paper uses Independent HN Accuracy to directly probe the cause/non-cause distinction targeted by preference supervision. The extended diagnostics below additionally separate construction-coupled and independently constructed hard-negative evaluations.

Constructed Hit@1 and Constructed Margin evaluate comparisons following the training-time counterfactual construction protocol. Independent HN Accuracy and Independent Margin use separately constructed hard negatives and therefore provide a less construction-coupled test. Counterfactual Consistency and Intervention Accuracy provide additional behavioral diagnostics under their corresponding evaluation protocols.

The distinction between the two headline hard-negative quantities is important: 93.45 is Constructed HN Hit@1, whereas 91.35 is Independent HN Accuracy. Balanced CPA-DPO improves Independent

Table 23: Paired fold-level comparison between final Balanced CPA-DPO and Vanilla-DPO on ECPE-CN.

Statistic	**Result**
Folds with Balanced CPA-DPO > Vanilla-DPO	10 / 10
Mean paired F1 difference	+1.81
Std. of paired differences	0.68
Paired t statistic	$t(9) = 8.42$
Cohen's d_z	2.66
p-value	$p < 0.001$
95% confidence interval	$[1.32,\ 2.30]$

HN Accuracy over Vanilla-DPO by

$$91.35 - 87.92 = +\mathbf{3.43} \tag{69}$$

percentage points. Because the independent hard negatives are constructed separately from the training-time counterfactual pool, this improvement is behavioral evidence that the learned discrimination extends beyond the exact preference examples used for optimization. These diagnostics do not establish that the model has learned a formal causal model or identified real-world causal effects.

D.11 Amplification Statistics

Using the final aggregate results, the LONG-minus-Overall gain amplification is

$$(75.46 - 70.42) - (85.43 - 83.62) = +\mathbf{3.23}, \tag{70}$$

and the Rebalanced-minus-Overall amplification is

$$(79.95 - 74.65) - (85.43 - 83.62) = +\mathbf{3.49}. \tag{71}$$

These point-estimate contrasts quantify the larger CPA-DPO advantage when the dominant locality regularity is less informative. Exact fold-level confidence intervals for these two amplification contrasts are not reconstructed here.

D.12 Fold-Level Statistical Analysis

The final matched Qwen2.5-32B comparison is also examined fold by fold. This analysis concerns the Overall F1 difference between Balanced CPA-DPO and Vanilla-DPO and is distinct from the amplification contrasts above.

For fold f, define

$$d_f = F1_f^{\mathrm{CPA}} - F1_f^{\mathrm{Vanilla}}, \qquad f = 1, \ldots, 10. \tag{72}$$

Across the ten folds,

$$\overline{d} = \mathbf{1.81}, \qquad s_d = \mathbf{0.68}. \tag{73}$$

Balanced CPA-DPO outperforms Vanilla-DPO on all ten folds. The paired analysis yields

$$t(9) = 8.42, \qquad p < 0.001, \tag{74}$$

with paired effect size

$$d_z = 2.66 \tag{75}$$

and a 95% confidence interval for the mean F1 improvement of

$$[1.32,\ 2.30]. \tag{76}$$

These statistics complement rather than replace the aggregate comparison reported in the main paper. They show that the mean improvement is not produced by only a small subset of folds and quantify the variability of the matched Balanced CPA-DPO–versus–Vanilla-DPO difference.

D.13 Summary of Supplementary Evidence

The supplementary analyses provide several consistency checks around the main-paper findings. First, threshold sensitivity shows that the default verifier configuration is not surrounded by an abrupt performance cliff. The final training data contain 6,178 chosen–rejected preference records: 96.08% match the reconstructed single-cause source pool and 3.92% are multi-cause-derived. The final data should therefore be described as cause-targeted preference supervision rather than universally one-cause-versus-one-trap comparisons.

Second, both the reconstructed source pool and final preference set remain strongly locality-skewed, while position-balanced construction yields near-chance position and construction-metadata probes. Human auditing further indicates that retained counterfactuals are generally plausible enough to provide non-trivial rejected alternatives and that the annotation-supported cause is usually clearly preferable.

Third, the final single-/multi-cause, backbone, RECCON, and extended hard-negative analyses broaden the evaluation beyond the headline ECPE-CN aggregate. The cross-backbone results preserve positive gains across all four evaluated backbones, and RECCON preserves the SFT $<$ Vanilla-DPO $<$ CPA-DPO ordering on both external conversational datasets. These transfer results are supplementary robustness/generalization evidence; RECCON is not used as direct evidence for the ECPE-CN locality-shortcut claim.

Finally, fold-level statistics show that the matched Overall improvement is consistent across all ten ECPE-CN folds, while the aggregate LONG and Rebalanced contrasts show substantially larger gains when locality is less informative. Together, these analyses support the narrower conclusion that CPA-DPO improves cause/non-cause discrimination and reduces reliance on the benchmark's dominant positional shortcut, while residual positional sensitivity and complete set-valued multi-cause modeling remain open limitations.

E Qualitative Analysis

The quantitative experiments in the main paper establish the primary performance and mechanism-related findings of Balanced CPA-DPO. This appendix complements those results with qualitative analyses of residual errors and successful corrections from the final position-balanced CPA-DPO system.

The purpose of these analyses is descriptive rather than causal. In particular, category frequencies are used to characterize recurrent prediction patterns and to examine whether they are consistent with the failure boundaries identified quantitatively. They should not be interpreted as decomposing the aggregate F1 gains into independent causal contributions.

E.1 Error Taxonomy

We randomly inspect $N = 100$ prediction errors made by the final position-balanced CPA-DPO system and assign each case to one primary error category. The categories are mutually exclusive for counting purposes: when a case exhibits multiple symptoms, it is assigned according to the dominant failure mode observed in the prediction.

Table 24 summarizes the resulting taxonomy.

Upstream emotion errors. The largest category consists of errors originating from the predicted emotion condition rather than from cause selection alone. These cases account for 35% of the sampled Balanced CPA-DPO errors. This observation is qualitatively consistent with the remaining performance headroom in the final Balanced run, where supplying the annotated emotion increases its F1 from 85.43 to 90.18.

The two quantities have different denominators and should not be interpreted as a direct numerical decomposition: the 35% statistic describes the composition of a sampled error set, whereas the 4.75-point difference compares aggregate F1 under predicted and gold emotion conditions. The qualitative conclusion is only that upstream emotion extraction remains an important source of residual end-to-end error.

Table 24: Qualitative taxonomy of 100 sampled Balanced CPA-DPO errors. Percentages are computed over the sampled error set rather than over all ECPE-CN test instances.

Error Category	Count	Share
Upstream emotion error	35	35%
Multi-cause / incomplete cause set	25	25%
Semantically ambiguous distractor	22	22%
Residual positional shortcut	10	10%
Formatting / index error	3	3%
Other	5	5%
Total	**100**	**100%**

Multi-cause and incomplete-set errors. Multi-cause or incomplete cause-set predictions account for 25% of the sampled errors. Typical cases occur when the model identifies one annotation-supported cause but omits another required cause, or otherwise fails to recover the complete annotated set.

This pattern is consistent with the slice analysis in Appendix D.7. Balanced CPA-DPO improves Vanilla-DPO by 2.05 F1 on single-cause instances but only 0.72 F1 on multi-cause instances.

The qualitative and quantitative results therefore identify the same scope boundary: the current CPA supervision is pair-level rather than complete set-level. Although a small fraction of retained preference records originates from multi-cause annotations, each constructed comparison still contrasts one annotation-supported cause against one counterfactual alternative. Complete multi-cause set prediction consequently remains more difficult.

Semantically ambiguous distractors. A further 22% of sampled errors involve distractor clauses that remain semantically plausible alternatives to the annotated cause. These clauses may share participants, events, affective content, or other contextual cues with the target emotion, making the distinction difficult even when simple locality cues are insufficient.

This category represents a remaining cause/non-cause discrimination challenge rather than a formatting or indexing failure. Its presence is also compatible with the hard-negative results: although the final Balanced run substantially improves Independent HN Accuracy, from 87.92 to 91.35 relative to Vanilla-DPO, the task is not solved perfectly.

Residual positional errors. Ten percent of the sampled errors remain compatible with a positional shortcut pattern. Operationally, these are cases in which the incorrect prediction favors a positionally attractive clause despite an annotation-supported alternative elsewhere in the document.

We use this category descriptively. A shortcut-compatible output does not by itself establish the internal decision mechanism used by the model. Nevertheless, the persistence of such cases is consistent with the controlled relocation results for the final Balanced system, which still exhibits a non-zero worst-case degradation of 2.72 F1 points. The qualitative analysis therefore supports the same bounded conclusion as the quantitative experiments: The Balanced CPA-DPO reduces positional shortcut dependence but does not eliminate positional sensitivity.

Formatting and index errors. Only 3% of sampled Balanced CPA-DPO errors are primarily attributable to formatting or clause-index failures. This is consistent with the system-level diagnostics in Appendix C.6, where 99.8% of outputs are valid JSON and 99.7% contain valid clause indices.

These statistics again use different denominators: the system-level rates are measured over generated outputs, whereas the 3% value is conditioned specifically on the sampled prediction errors. Accordingly, the comparison is used only to show that malformed output is not a dominant qualitative failure mode.

E.2 Successful Correction Taxonomy

We next inspect $N = 40$ cases in which Vanilla-DPO produces an incorrect prediction but CPA-DPO produces the annotation-consistent prediction under the matched evaluation protocol. The goal is to

Table 25: Qualitative taxonomy of 40 sampled cases corrected by final position-balanced CPA-DPO relative to Vanilla-DPO. The distribution describes the sampled correction set and should not be interpreted as the distribution of all ECPE-CN examples.

Correction Type	**Count**	**Share**
Long-distance / locality-conflict	17	42.5%
Nearby plausible distractor rejected	12	30.0%
Emotionally salient annotation-inconsistent distractor	8	20.0%
Other	3	7.5%
Total	**40**	**100.0%**

characterize the types of errors that are corrected, rather than to estimate their prevalence over the complete test set.

Table 25 reports the resulting categories.

Long-distance and locality-conflict corrections. The largest category contains long-distance or locality-conflict cases, accounting for 17 of the 40 sampled corrections (42.5%). In these examples, the annotation-supported cause is not the most positionally attractive candidate, and the Vanilla-DPO prediction is incorrect while Balanced CPA-DPO recovers the annotated cause.

This pattern qualitatively aligns with the substantially larger quantitative improvement on LONG examples: The final Balanced system improves Vanilla-DPO from 70.42 to 75.46 LONG F1, a gain of 5.04 points, compared with a 1.81-point gain on the complete ECPE-CN benchmark in that final run.

The 42.5% qualitative share does not establish that these cases account for 42.5% of the aggregate LONG improvement. It only shows that long-distance and locality-conflict examples occur frequently within the sampled set of successful corrections.

Rejection of nearby plausible distractors. Twelve cases (30.0%) involve a nearby clause that is superficially plausible under the document context but is not the annotated cause. Vanilla-DPO selects the competing clause, whereas Balanced CPA-DPO rejects it in favor of the annotation-supported cause.

These cases qualitatively match the intended role of counterfactual preference supervision: the rejected alternative remains competitive under coarse contextual or positional cues, so successful prediction requires a finer distinction between the annotation-supported cause and a plausible competitor. The pattern is also consistent with the final improvement in Independent HN Accuracy from 87.92 to 91.35.

Emotionally salient annotation-inconsistent distractors. Eight sampled corrections (20.0%) involve clauses that are emotionally salient or topically related to the target emotion but are inconsistent with the annotated cause relation. We deliberately use the term *annotation-inconsistent distractor* rather than asserting that such clauses are objectively non-causal.

These cases illustrate a distinction central to the preference construction: semantic or affective relevance alone is not sufficient for selecting the annotation-supported cause. Balanced CPA-DPO is trained to prefer the annotation-supported output over plausible alternatives that would otherwise remain competitive.

Other corrections. The remaining three cases (7.5%) do not fit cleanly into the preceding categories. They include heterogeneous corrections for which no single positional, semantic, or salience-based pattern dominates. We retain this residual category rather than forcing all examples into the mechanism-oriented taxonomy.

E.3 Relationship to the Quantitative Evidence

The two qualitative analyses are consistent with, but do not replace, the quantitative evidence in the main paper.

First, the high frequency of upstream emotion errors and the smaller multi-cause gain identify clear remaining system boundaries. Second, the presence of residual positional errors is consistent with the non-zero degradation under controlled relocation and therefore reinforces the claim of reduced, rather than eliminated, positional shortcut dependence. Third, long-distance/locality-conflict corrections and rejection of nearby plausible distractors qualitatively align with the larger LONG gain and the improvement on independent hard negatives.

These observations should be interpreted as triangulation across evaluation views rather than as a causal decomposition of performance. The main conclusions of the paper continue to rest on the controlled quantitative comparisons reported in the main text and Appendix D.