SELECTIVE EVIDENCE LOSS: TASK-RELEVANT MISSINGNESS IN MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world multimodal systems can lose information selectively through content filtering and privacy protection. Existing incomplete multimodal learning mainly characterizes missingness by its amount and structure, leaving the task relevance of missing content underexplored. We introduce Selective Evidence Loss (SEL) to study task-relevant missingness under controlled deletion, and operationalize task relevance in multimodal sentiment analysis using contextual affective intensity. SEL causes greater performance degradation than random deletion under matched deletion amounts, and the gap persists after matching affected utterances and within-utterance deletion structure. This vulnerability is not consistently resolved by mechanism-matched training. We propose Selective Relational Distillation and Preservation (SRDP), which provides relational supervision for affected samples while preserving model behavior on unaffected samples. SRDP produces a single deployment checkpoint and requires only the corrupted multimodal input at inference. In the main evaluation, SRDP reduces affected-sample MAE for all ten models, with a model-average reduction of 0.012. These results show that robustness depends on what evidence is lost, highlighting task relevance as a key dimension of missing-information robustness.

1 INTRODUCTION

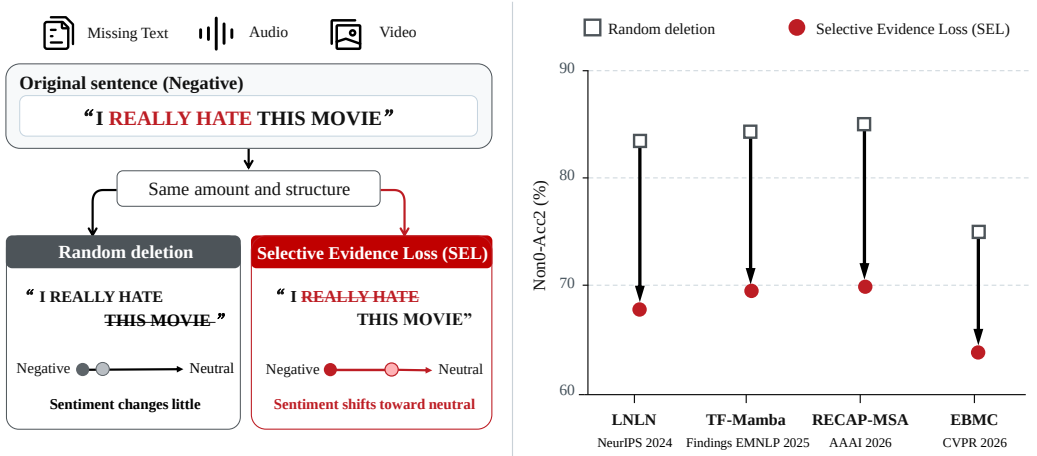
Multimodal sentiment analysis integrates complementary evidence from text, audio, and visual signals to understand affective expressions (Sun & Tian, 2025; He et al., 2026b; Zhang et al., 2023). In real-world systems, parts of this evidence usually become unavailable because of data loss, privacy-preserving anonymization, or content-dependent filtering (Wu et al., 2026; Loiseau et al., 2025; Mei et al., 2026). Such information loss can be selective, since anonymization and filtering decisions can depend on the content itself (Mei et al., 2026; Loiseau et al., 2025). The consequence of such selective loss depends on which evidence is removed, even when the overall amount of missing information is unchanged.

Consider the sentence **I really hate this movie.** As illustrated in Figure 1 (a), removing **this movie** and removing **really hate** can produce the same missing amount and span structure, yet the two deletions remove evidence with different relevance to sentiment prediction. The former preserves the explicit affective expression, whereas the latter removes it. We define task relevance as the contribution of content to the target prediction. When missing-content selection depends on this relevance, we refer to the resulting process as task-relevant missingness.

Existing evaluations on missing multimodal inputs commonly vary the unavailable modality, missing amount, position, or corruption pattern (Li et al., 2024a; Zhang et al., 2024; Li et al., 2025; Zhuang et al., 2026; He et al., 2026a). These settings characterize how much information is missing and how the missingness is structured, but they do not control the task relevance of the removed content. Consequently, two inputs with comparable missing amount and structure can still retain very different amounts of task-relevant evidence. This motivates a controlled question: when missing amount and deletion structure are held comparable, does task-relevant missingness systematically change model performance?

To study this question, we introduce **Selective Evidence Loss (SEL)**, which selectively removes textual evidence according to its task relevance. In multimodal sentiment analysis, we operationalize

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.



(a) Same structure Different task relevance **(b) Same structure Different accuracy**

Figure 1: **Selective evidence loss under matched deletion structure.** (a) illustrates two deletions with the same deleted-word count and span structure but different task relevance: Random deletion preserves the explicit affective expression, while SEL removes it. (b) presents Non0-Acc2 on the MOSI AER60 common-support test subset for four representative models, averaged over three training seeds. Each arrow connects Structure-Matched Random and SEL for the same model, showing consistently larger performance degradation when more task-relevant evidence is removed.

this relevance using contextual affective intensity. Figure 1 (b) illustrates the effect on four representative models under the structure-matched setting at MOSI AER60. Across the full evaluation of ten models, two datasets, and three severity levels, SEL consistently produces greater degradation under both deletion-amount-matched and structure-matched comparisons. This gap shows that task relevance affects robustness even when deletion amount and structure are matched.

These findings establish task relevance as an additional dimension of missing-information robustness beyond missing amount and structure. Evaluations based only on random deletion can overlook vulnerability to selective loss of task-relevant evidence. They also raise a training challenge. Mechanism-matched training on SEL-corrupted inputs yields inconsistent gains across models and conditions. Under SEL, only a subset of samples is actually altered by selective deletion, creating different adaptation needs for affected and unaffected samples. Affected samples need targeted supervision, while unaffected samples should preserve their established model behavior.

To address these two training needs, we introduce Selective Relational Distillation and Preservation (SRDP). For affected samples, SRDP transfers pairwise relational structure from a frozen reference model, providing structural guidance without imposing exact pointwise agreement with the reference. For unaffected samples, SRDP constrains unnecessary drift from the Native model, preserving behavior already learned from unchanged inputs. Together, these components provide sample-conditional adaptation to selective evidence loss. The resulting model is deployed as a single checkpoint with the same architecture and parameter count as the underlying base model, requiring only the corrupted multimodal input at test time.

Our contributions are summarized below:

- We introduce task-relevant missingness as a dimension of missing-information robustness and formulate Selective Evidence Loss (SEL) to study it under controlled deletion.
- We demonstrate that removing more task-relevant content consistently leads to greater performance degradation under matched deletion amount and structure.
- We propose Selective Relational Distillation and Preservation (SRDP), a sample-conditional training strategy that provides relational supervision for affected samples while preserving model behavior on unaffected samples.

2 Related Work

2.1 Multimodal Sentiment Analysis

Multimodal sentiment analysis predicts sentiment from linguistic, acoustic, and visual signals. Existing methods model cross-modal information through attention and feature interaction (Tsai et al., 2019; Wang et al., 2023), representation constraints and modality-specific supervision (Hazarika et al., 2020; Yu et al., 2021; Han et al., 2021), or language-guided multimodal learning (Zhang et al., 2023; Sun & Tian, 2025).

These methods differ in how information from individual modalities is represented, aligned, and integrated into the final prediction. Such differences can also affect how a model responds when informative content becomes unavailable. We evaluate task-relevant missingness with different multimodal learning strategies, rather than restricting the analysis to a single fusion architecture.

2.2 Incomplete Multimodal Learning

Incomplete multimodal learning studies prediction when part of a multimodal input is unavailable (Wu et al., 2026). Current approaches can be broadly grouped into recovery-based and recovery-free methods. Recovery-based methods reconstruct missing inputs or representations through tensor completion, language guidance, modality complementation, denoising, or uncertainty-aware alignment (Wu & Fan, 2024; Zeng et al., 2022; Li et al., 2025; Zhuang et al., 2026). Recovery-free methods improve robustness through distillation, knowledge transfer, hierarchical learning, or adaptive modality collaboration (Li et al., 2024a;c;b; He et al., 2026b).

Existing work mainly studies how to improve prediction under predefined missing-input conditions (Zhang et al., 2024; Li et al., 2024b). These conditions are commonly specified by the unavailable modality, missing amount, position, or corruption pattern (Yuan et al., 2021; Zhang et al., 2024; Li et al., 2025). Related work has also studied downstream learning directly in the presence of missing values (Xiao & Fan, 2024). We complement these settings by introducing task relevance into the missing-content selection and evaluating its effect under matched deletion controls.

2.3 Missingness Mechanisms

Missing-data research has long shown that the mechanism determining what becomes unavailable can affect subsequent analysis (Rubin, 1976; Seaman et al., 2013; Little & Rubin, 2019). Missingness depends on properties of the data itself, making the selection process an important part of how incomplete observations are formed. Recent surveys further extend this perspective to modern machine-learning settings and complex data types, including multimodal data (Fan, 2026). This perspective is especially relevant to content filtering and anonymization, where the probability of removal can depend on the content being processed.

Our work brings this mechanism perspective to incomplete multimodal learning. We treat missing-content selection as an explicit factor in robustness evaluation and study a setting in which the selection process is guided by task relevance. This connects content-dependent missingness with controlled robustness analysis in multimodal sentiment prediction.

3 Selective Evidence Loss (SEL)

Selective Evidence Loss (SEL) provides a controlled setting for studying task-relevant missingness. We distinguish how much information is removed from what content is selected for removal.

3.1 Formalizing Missing-Content Selection

We study multimodal sentiment analysis under partial loss of textual evidence. A supervised sample is $z = (x^c, y)$, where $x^c = (T, A, V)$ contains textual, acoustic, and visual inputs, y is the sentiment label, and $T = (w_1, \ldots, w_n)$ is the surface-word sequence.

A missingness mechanism selects $D \subseteq \{1, \ldots, n\}$ and produces

$$x^m(D) = \big(\widetilde{T}(D), A, V\big), \qquad \widetilde{T}(D) = (w_i \mid i \notin D)\,, \tag{1}$$

where the remaining words preserve their original order. We focus on textual deletion because transcript content can be selectively lost through filtering or transcription errors, while holding audio and visual inputs fixed isolates the effect of textual content selection.

The deletion amount $|D|$ specifies how much text is removed. Let E denote a textual evidence unit and $r(E;T)$ its contribution to the target task. Deletion sets with the same $|D|$ can remove evidence with different task relevance. For example, in a MOSI utterance, SEL removes "VERY VERY EXCITED" from "I WENT TO THIS MOVIE VERY VERY EXCITED", while Structure-Matched Random removes "I WENT TO". Both remove one three-word span, but only SEL removes the explicit affective expression. We characterize missingness by both deletion amount and the rule used to select the missing content.

3.2 Operationalizing Task Relevance

For multimodal sentiment analysis, since $r(E;T)$ is not directly observed, we use contextual affective intensity as an operational proxy for task relevance. Candidate expressions are scored in context and resolved into non-overlapping evidence units,

$$\mathcal{E}(T) = \big\{(E_j, e_j)\big\}_{j=1}^{J}, \qquad e_j \in [0,1], \tag{2}$$

where E_j is a retained affective expression and e_j is its contextual affective intensity. SEL removes evidence units in descending order of e_j, preferentially targeting more sentiment-relevant textual evidence. Human alignment of this proxy is evaluated in Section 5.2, while the complete extraction and scoring procedure is reported in Appendix A.

We define Affective Evidence Removal (AER) over the total eligible evidence-word capacity of each split. AER20, AER40, and AER60 remove 20%, 40%, and 60% of this capacity, respectively, with nested deletion sets across severity levels.

3.3 Controlled Random Comparisons

We compare SEL with two controls to isolate the effect of missing-content selection. Budget-Matched Random matches the overall deletion amount while allowing affected utterances and deletion structures to vary. Structure-Matched Random further matches the affected utterances and within-utterance deletion structure on a common support. These controls test whether model performance depends on what content is removed after controlling for deletion amount and, in the stricter comparison, deletion structure. Detailed construction procedures are provided in Appendix A.

4 Method

We first train a mechanism-matched baseline, called Native, on SEL-corrupted data using the original objective of each base model. Selective Relational Distillation and Preservation (SRDP) then provides relational supervision for affected samples and preserves Native behavior on unaffected samples. Figure 2 summarizes the training procedure.

4.1 Training Setup

For sample i, let $x_i^m = (\widetilde{T}_i, A_i, V_i)$ and $x_i^c = (T_i, A_i, V_i)$ denote its corrupted and clean views. A training-only indicator $d_i \in \{0,1\}$ records whether SEL deletes at least one surface word. For mini-batch $\mathcal{B}$, we define

$$\mathcal{B}_a = \{i \in \mathcal{B} : d_i = 1\}, \qquad \mathcal{B}_u = \{i \in \mathcal{B} : d_i = 0\}, \tag{3}$$

for affected and unaffected samples, respectively.

For base model f_θ, let $f_\theta(x) = [h_\theta(x), p_\theta(x)]$, where $h_\theta(x) \in \mathbb{R}^{d_h}$ is the fused representation and $p_\theta(x) \in \mathbb{R}$ is the sentiment prediction.

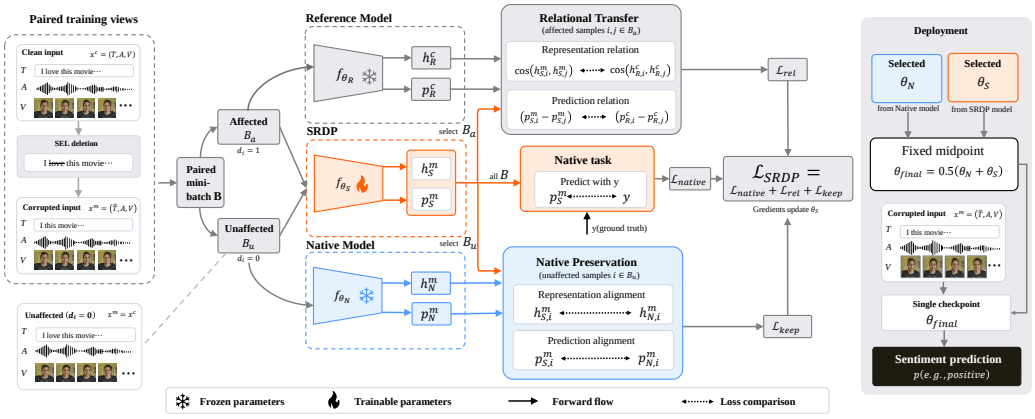


Figure 2: **Selective Relational Distillation and Preservation.** SRDP applies relational supervision to affected samples and Native-model preservation to unaffected samples. The resulting single checkpoint uses only corrupted multimodal input at inference.

We denote the selected Native checkpoint by θ_N and the trainable SRDP model by θ_S. During SRDP training, θ_N is frozen and serves as the preservation anchor. A separately trained and validation-selected frozen reference θ_R provides relational targets for the same model, dataset, severity level, and seed. In the implemented SRDP, θ_R processes the clean view; Section 5 separately examines this reference choice. Native and SRDP otherwise share the same base-model architecture and native training configuration.

4.2 Relational Transfer

For affected sample i, the trainable model processes x_i^m and the frozen reference processes x_i^c:

$$(h_{S,i}^m, p_{S,i}^m) = f_{\theta_S}(x_i^m), \qquad (h_{R,i}^c, p_{R,i}^c) = f_{\theta_R}(x_i^c).$$

Pairwise supervision constrains relative organization without requiring exact pointwise agreement with the reference, consistent with recent correlation-aware distillation methods for incomplete multimodal learning (Li et al., 2024c). For each $i \neq j$ in $\mathcal{B}_a$, we define

$$\begin{aligned} s_{ij}^m &= \cos(h_{S,i}^m, h_{S,j}^m), \quad & s_{ij}^c &= \cos(h_{R,i}^c, h_{R,j}^c), \\ \delta_{ij}^m &= p_{S,i}^m - p_{S,j}^m, & \delta_{ij}^c &= p_{R,i}^c - p_{R,j}^c. \end{aligned} \tag{4}$$

Let $\mathcal{P}_a = \{(i,j) : i,j \in \mathcal{B}_a,\ i \neq j\}$. Using Smooth L1 loss $\rho(\cdot,\cdot)$ and stop-gradient $\mathrm{sg}(\cdot)$,

$$\begin{aligned} \mathcal{L}_{\mathrm{rel}} = \frac{1}{|\mathcal{P}_a|} \sum_{(i,j)\in\mathcal{P}_a} \Big[& \lambda_{\mathrm{corr}} \rho\big(s_{ij}^m, \mathrm{sg}(s_{ij}^c)\big) \\ & + \lambda_{\mathrm{dist}} \rho\big(\delta_{ij}^m, \mathrm{sg}(\delta_{ij}^c)\big) \Big]. \end{aligned} \tag{5}$$

This objective aligns corrupted-view representation geometry and relative prediction differences with the frozen reference. The native task objective continues to anchor absolute predictions. When $|\mathcal{B}_a| < 2$, we set $\mathcal{L}_{\mathrm{rel}} = 0$.

4.3 Native Preservation

Updates driven by affected samples can also alter model behavior on unchanged inputs, echoing the broader problem of preserving previously learned behavior during adaptation (Li & Hoiem, 2016). For $i \in \mathcal{B}_u$, where $x_i^m = x_i^c$, we preserve the frozen Native prediction and representation:

$$\begin{aligned} \mathcal{L}_{\mathrm{keep}} = \frac{1}{|\mathcal{B}_u|} \sum_{i\in\mathcal{B}_u} \Big[& \lambda_{\mathrm{pred}} \rho(p_{\theta_S}(x_i^m), \mathrm{sg}(p_{\theta_N}(x_i^m))) \\ & + \lambda_{\mathrm{feat}} \Big(1 - \cos\big(h_{\theta_S}(x_i^m), \mathrm{sg}(h_{\theta_N}(x_i^m))\big)\Big) \Big]. \end{aligned} \tag{6}$$

This discourages unnecessary drift on unaffected samples. When $|\mathcal{B}_u| = 0$, we set $\mathcal{L}_{\mathrm{keep}} = 0$.

4.4 SRDP OBJECTIVE

The complete objective is

$$\mathcal{L}_{\mathrm{SRDP}} = \mathcal{L}_{\mathrm{native}} + \mathcal{L}_{\mathrm{rel}} + \mathcal{L}_{\mathrm{keep}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{native}}$ is the original base-model objective on the corrupted mini-batch. The auxiliary coefficients are fixed across all evaluated configurations; their values are reported in Appendix B.

After selecting θ_N and θ_S independently on SEL-corrupted validation data, we form the deployment checkpoint

$$\theta_{\mathrm{final}} = \tfrac{1}{2}\big(\theta_N + \theta_S\big). \tag{8}$$

This parameter-space integration follows the broader observation that averaging compatible model weights can improve performance without increasing inference cost (Wortsman et al., 2022). The interpolation coefficient is fixed at 0.5 for all configurations. The resulting checkpoint has the same architecture and parameter count as the base model. The reference and Native branches are used only during training; inference requires one forward pass of θ_{final} on $(\widetilde{T}, A, V)$.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate ten models on the unaligned MOSI and MOSEI splits at three Affective Evidence Removal (AER) levels: AER20, AER40, and AER60. The evaluated models include six conventional multimodal sentiment models and four models designed for incomplete multimodal inputs. For each model, severity level, and seed, SEL and both random controls are evaluated using the same Budget-Matched Random-trained checkpoint, while SRDP is compared with its paired Native model trained on SEL-corrupted data.

Non0-Acc2 is the primary metric for the SEL robustness comparisons. Affected MAE is the primary outcome for SRDP because it measures prediction error on samples whose text is changed by SEL while retaining zero-valued sentiment labels. We additionally report other metrics and results on the Full and Unaffected subsets where applicable. Complete dataset statistics, metric definitions, model configurations, checkpoint selection, and aggregation procedures are provided in Appendix A.6 and Appendix B.2.

5.2 VALIDATING THE TASK-RELEVANCE PROXY

We conduct a blinded three-annotator audit to assess whether SEL removes more sentiment-relevant content than Structure-Matched Random. After quality control, 315 deletion pairs yield 932 valid judgments. Among them, 82.4% favor SEL, 11.9% indicate a tie, and 5.7% favor Structure-Matched Random. The mean oriented score is 1.432, with a 95% item-clustered bootstrap interval of $[1.352, 1.509]$.

Agreement on the exact ordinal strength is modest (Krippendorff's $\alpha = 0.278$), while the aggregate direction strongly favors SEL. We interpret these results as human support for contextual affective intensity as a sentiment-relevance proxy, rather than a direct measure of each expression's task contribution. The full annotation protocol, quality-control procedure, and results by dataset and severity are reported in Appendix A.5.

5.3 ROBUSTNESS UNDER TASK-RELEVANT MISSINGNESS

We test whether model robustness depends on which content is removed after progressively controlling for deletion amount and deletion structure.

Matched deletion amount. SEL and Budget-Matched Random remove the same total number of words within each dataset split and severity, while the affected utterances and deletion structures may differ. Across three-seed means, SEL lowers Non0-Acc2 in all 60 model–dataset–severity conditions. Figure 3 illustrates Non0-Acc2 in panels a and c and MAE in panels b and d. The model-average Non0-Acc2 gaps at AER20, AER40, and AER60 are 2.8, 3.1, and 4.7 percentage

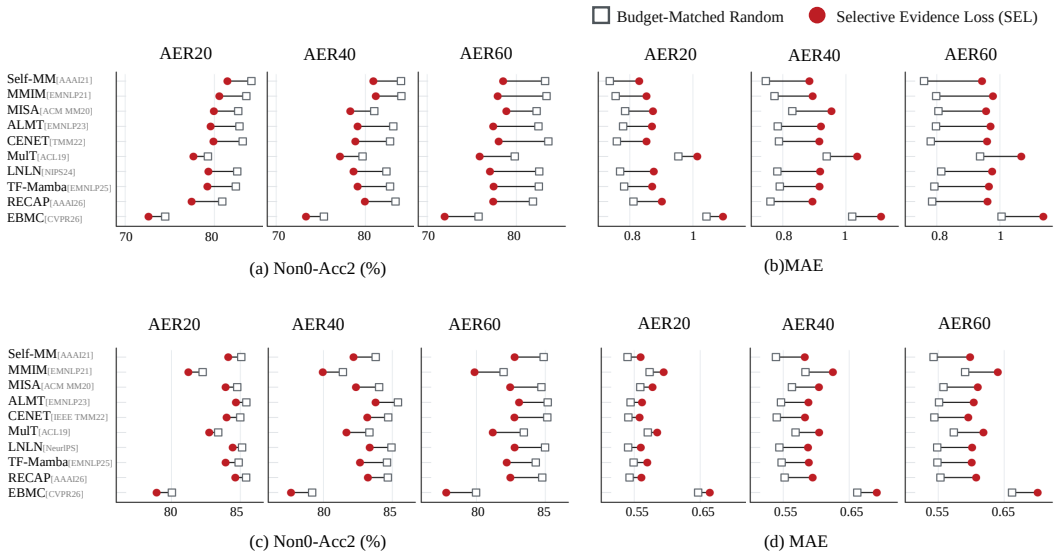


Figure 3: **SEL versus Budget-Matched Random on the full test split.** Panels a and b illustrate Non0-Acc2 and MAE on MOSI. Panels c and d present Non0-Acc2 and MAE on MOSEI. All results are three-seed means. Smaller MAE indicates better performance.

points on MOSI and 0.9, 1.6, and 2.2 on MOSEI. These results show that equal overall deletion amounts can still produce systematically different robustness when different content is removed. Exact model-level Non0-Acc2 scores appear in Appendix Table A4.

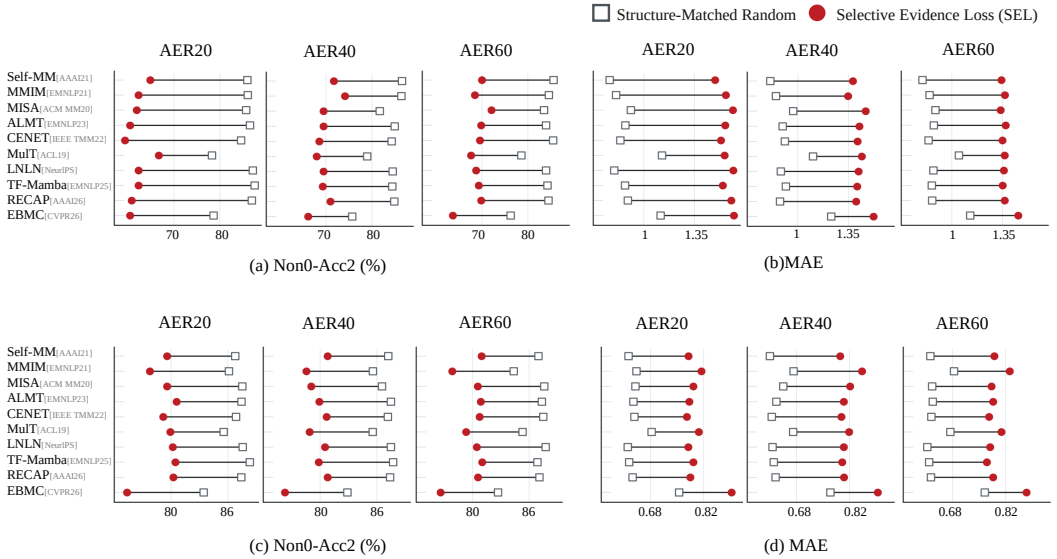


Figure 4: **SEL versus Structure-Matched Random on the condition-specific common support.** Panels a and b provide Non0-Acc2 and MAE on MOSI. Panels c and d report Non0-Acc2 and MAE on MOSEI. All results are three-seed means. Smaller MAE indicates better performance.

Matched deletion structure. We next compare SEL with Structure-Matched Random on the condition-specific common support of affected utterances admitting a valid alternative deletion. For each matched utterance, the two conditions share its identity, deleted-word count, number of spans,

and span lengths, differing only in the deleted content. Across three-seed means, SEL again lowers Non0-Acc2 in all 60 conditions. Figure 4 presents Non0-Acc2 in panels a and c and MAE in panels b and d. The model-average Non0-Acc2 gaps at AER20, AER40, and AER60 are 22.6, 13.2, and 14.2 percentage points on MOSI and 7.4, 7.0, and 6.5 on MOSEI. These results show that task relevance affects robustness beyond deletion amount and structure. Exact model-level scores appear in Appendix Table A5. The gap magnitudes from the two comparisons are not directly comparable because they use different test supports.

Table 1: **Native versus SRDP on MOSEI at AER60.** Results are averaged over three paired training seeds. Parentheses report SRDP−Native for MAE; lower MAE is better.

Model	Method	Non0-Acc2 ↑	Non0-F1 ↑	MAE ↓	Corr ↑	Affected MAE ↓	Unaffected MAE ↓
Self-MM	Native	81.5	81.4	0.602	0.662	0.775	0.519
	SRDP	**82.1**	**82.2**	**0.586 (−0.016)**	**0.680**	**0.758 (−0.017)**	**0.503 (−0.016)**
MMIM	Native	80.0	79.6	0.626	0.638	0.810	0.538
	SRDP	**81.3**	**81.0**	**0.600 (−0.026)**	**0.669**	**0.774 (−0.036)**	**0.517 (−0.021)**
MISA	Native	81.3	81.3	0.618	0.666	0.793	0.535
	SRDP	**82.4**	**82.4**	**0.607 (−0.011)**	**0.673**	**0.779 (−0.015)**	**0.526 (−0.009)**
ALMT	Native	82.5	82.5	0.604	0.678	0.773	0.523
	SRDP	**82.9**	**82.8**	**0.599 (−0.005)**	**0.684**	**0.762 (−0.011)**	**0.521 (−0.002)**
CENET	Native	82.8	82.8	0.593	0.679	0.770	0.509
	SRDP	**83.0**	**83.0**	**0.591 (−0.003)**	**0.683**	**0.765 (−0.005)**	**0.508 (−0.002)**
MulT	Native	80.1	80.1	0.621	0.646	0.803	0.533
	SRDP	**81.4**	**81.4**	**0.614 (−0.007)**	**0.651**	**0.797 (−0.007)**	**0.526 (−0.007)**
LNLN	Native	81.9	82.0	0.610	0.676	0.776	0.531
	SRDP	**82.1**	**82.1**	**0.609 (−0.001)**	**0.677**	**0.774 (−0.002)**	**0.530 (−0.001)**
TF-Mamba	Native	82.0	82.0	0.608	0.668	0.764	0.534
	SRDP	**82.3**	**82.2**	**0.598 (−0.010)**	**0.678**	**0.758 (−0.006)**	**0.522 (−0.013)**
RECAP-MSA	Native	75.6	**74.9**	0.680	0.516	0.874	0.587
	SRDP	**75.8**	74.9	**0.674 (−0.006)**	**0.525**	**0.868 (−0.006)**	**0.581 (−0.006)**
EBMC	Native	76.9	76.8	0.705	0.568	0.873	0.625
	SRDP	**77.8**	**77.7**	**0.685 (−0.020)**	**0.581**	**0.860 (−0.013)**	**0.602 (−0.023)**

5.4 Limits of Mechanism-Matched Training

We compare Budget-Matched Random-trained and SEL-trained checkpoints on the same full SEL test split. Mechanism-matched training lowers mean MAE in 17 of 60 model–dataset–severity conditions and raises mean Non0-Acc2 in 29 of 60. These mixed gains indicate that exposure to SEL corruption alone is not a consistently reliable adaptation strategy, motivating sample-conditional objectives for affected and unaffected samples. Complete model-level results are provided in Appendix Figure A1.

5.5 SRDP under Task-Relevant Missingness

Affected MAE is the primary outcome for evaluating SRDP. Table 1 provides the main comparison between Native and SRDP. On MOSEI at AER60, SRDP lowers the three-seed mean Affected MAE for all ten models, with a model-average reduction of 0.012. Full and Unaffected MAE also decrease for all ten models, each by 0.010 on average. Secondary metrics show condition-dependent trade-offs. For example, at MOSEI AER40, Self-MM improves Full MAE while Full Non0-Acc2 decreases. Complete metric profiles are reported in Appendix Figure A3, and seed-specific intervals for the ten-model comparison are shown in Appendix Figure A2.

5.6 Component Analysis and Controls

Figure 5 compares Native, the selected SRDP student, and the fixed midpoint across all ten models on MOSEI AER60. The midpoint lowers mean MAE relative to Native for every model and generally improves upon the selected student across the Full, Affected, and Unaffected subsets.

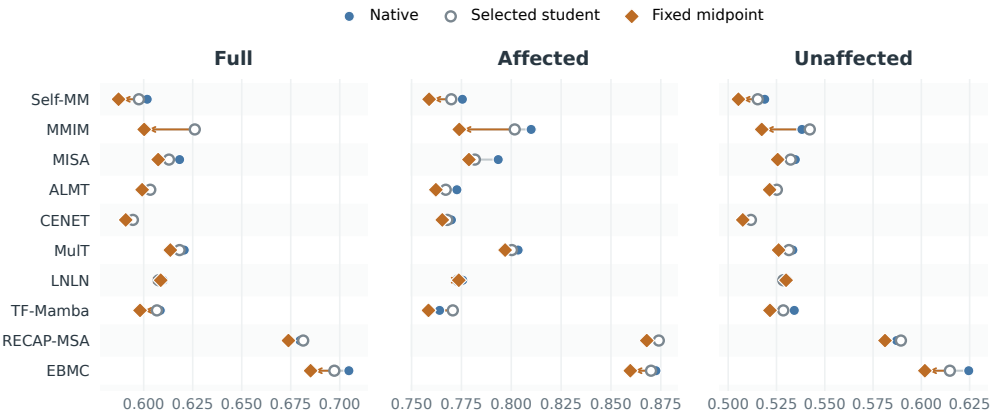


Figure 5: **Fixed parameter midpoint on MOSEI at AER60.** Each marker shows the three-seed mean MAE for one model. Native, the selected SRDP student, and the fixed midpoint are compared on the Full, Affected, and Unaffected subsets. Lower values are better.

Table 2 contains the main training-strategy controls and component ablations on Self-MM at MOSEI AER60. SRDP achieves a larger Affected MAE reduction than clean–corrupted joint training and pointwise distillation. Shuffling pairwise correspondence or removing Native preservation increases validation Affected MAE, supporting both relational supervision and preservation. Additional checkpoint, uncertainty, and reference-view analyses are provided in Appendix C.

Table 2: **Training controls and component analysis on Self-MM at MOSEI AER60.** Three-seed mean $\pm$ standard deviation; test controls are relative to Native and validation ablations to full SRDP.

Test controls		*Validation ablations*	
Clean+Corrupted Joint	-0.0039 ± 0.0098	Shuffled Correspondence	$+0.0102 \pm 0.0014$
Pointwise Distillation	-0.0130 ± 0.0024	Without Native Preservation	$+0.0168 \pm 0.0063$
SRDP	-0.0167 ± 0.0082	All-Sample Relations w/o Preservation	$+0.0217 \pm 0.0093$
		Without Representation Relations	$+0.0048 \pm 0.0067$

6 Limitations

Our study is limited to controlled textual evidence loss in multimodal sentiment analysis, with contextual affective intensity as an operational proxy for task relevance; other languages, modalities, and naturally occurring missingness remain to be studied. SRDP requires paired clean and corrupted views during training, while inference uses only the corrupted multimodal input.

7 Conclusion

We study task-relevant missingness as an explicit dimension of robustness in multimodal sentiment analysis. SEL shows that removing more task-relevant content causes greater performance degradation even when deletion amount and within-utterance structure are controlled. A blinded human audit further supports contextual affective intensity as an operational proxy for task relevance. We also introduce SRDP, which provides relational supervision for affected samples while preserving model behavior on unaffected samples. On MOSEI at AER60, SRDP reduces affected-sample MAE across all ten evaluated models, supporting task relevance as an important factor in both robustness evaluation and training under missing information.

AI USE STATEMENT

Generative AI tools were used to assist language editing, figure prototyping, and code drafting. All AI-assisted text, visual content, code, calculations, and scientific claims were reviewed by the authors against the underlying methods and experimental records. The authors take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

The benchmark construction, deletion controls, SRDP objectives, training protocol, checkpoint selection, and aggregation procedures are documented in the main text and appendix. The appendix further provides artifact audits, exact benchmark results, implementation details, uncertainty analyses, and additional controls needed to reproduce the reported experiments.

REFERENCES

Jicong Fan. An interdisciplinary and cross-task review on missing data imputation. *Foundations and Trends in Signal Processing*, 20(3):185–317, 2026.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM International Conference on Multimedia (ACM MM)*, 2020.

Kang He, Yuzhe Ding, Rao Fu, Yukang Feng, Kaipeng Zhang, Yiming Liu, Fei Li, Chong Teng, and Donghong Ji. Uncertainty-calibrated elastic alignment for multimodal sentiment analysis with missing modalities. In *Findings of the Association for Computational Linguistics (Findings of ACL)*, 2026a.

Kang He, Yuzhe Ding, Xinrong Wang, Fei Li, Chong Teng, and Donghong Ji. Enhance-then-balance modality collaboration for robust multimodal sentiment analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026b.

Huiting Huang, Tieliang Gong, Kai He, Wen Wen, Weizhan Zhang, and Mengling Feng. Recovering coherent affective patterns: Addressing modality missing in multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 21957–21965, 2026. doi: 10.1609/aaai.v40i26.39349.

Mingcheng Li, Dingkang Yang, Yuxuan Lei, Shunli Wang, Shuaibing Wang, Liuzhen Su, Kun Yang, Yuzheng Wang, Mingyang Sun, and Lihua Zhang. A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2024a.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, Mingyang Sun, Ziyun Qian, Dongliang Kou, and Lihua Zhang. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024c.

Xiang Li, Xianfu Cheng, Dezhuang Miao, Xiaoming Zhang, and Zhoujun Li. TF-mamba: Text-enhanced fusion mamba with missing modalities for robust multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics (Findings of EMNLP)*, 2025.

Zhizhong Li and Derek Hoiem. Learning without forgetting. In *European Conference on Computer Vision*, pp. 614–629. Springer, 2016.

Roderick J. A. Little and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2019.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. Tau-Eval: A unified evaluation framework for useful and private text anonymization. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP System Demonstrations)*, 2025.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, Ruibin Yuan, and Xueqi Cheng. HiddenGuard: Fine-grained safe generation with specialized representation router. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Donald B. Rubin. Inference and missing data. *Biometrika*, 63:581–592, 1976.

Shaun Seaman, John Galati, Dan Jackson, and John Carlin. What is meant by "missing at random"? *Statistical Science*, 28:257–268, 2013.

Kaiwei Sun and Mi Tian. Sequential fusion of text-close and text-far representations for multimodal sentiment analysis. In *International Conference on Computational Linguistics (COLING)*, 2025.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. Cross-modal enhancement network for multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 25:4909–4921, 2023.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: Averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 23965–23998. PMLR, 2022.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026.

Tongle Wu and Jicong Fan. Smooth tensor product for tensor completion. *IEEE Transactions on Image Processing*, 33:6483–6496, 2024. doi: 10.1109/TIP.2024.3489272.

Feng Xiao and Jicong Fan. Unsupervised anomaly detection in the presence of missing values. *Advances in Neural Information Processing Systems*, 37:138130–138162, 2024.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *ACM International Conference on Multimedia (ACM MM)*, 2021.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Haoyu Zhang, Yu Wang, Guanghao Yin, Kejun Liu, Yuanyuan Liu, and Tianshu Yu. Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. TMDC: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

A SEL CONSTRUCTION, VALIDATION, AND COMPLETE RESULTS

This section describes SEL construction, the matched random controls, human validation of the relevance proxy, benchmark statistics, and the complete robustness and mechanism-matched training results.

A.1 CANDIDATE EXTRACTION

Candidate extraction operates on non-punctuation surface words obtained from a Stanza part-of-speech and dependency parse. Adjective, adverb, and verb heads seed candidate expressions, which are expanded using fixed local rules for negation, modifiers, and verb particles. Singleton *more* and *most* are excluded. Selected surface words are deleted before model-specific subword tokenization, after which the corrupted transcript is tokenized and encoded anew.

A.2 AFFECTIVE SCORING

Each candidate expression is evaluated in context to determine eligibility and contextual affective intensity $e_j \in [0, 1]$. Exact NRC-VAD v2.1 matches provide lexical affective information when the absolute valence reaches the fixed threshold of 0.5, with arousal mapped to $[0, 1]$ as intensity. For candidates without an exact expression-level match, the procedure uses direct head or phrasal evidence when available; otherwise, semantic retrieval and a fixed contextual language-model procedure determine affective identity, polarity, and intensity. Neutral, ambiguous, or failed candidates are excluded. Polarity and confidence are retained for auditing but do not affect deletion priority after eligibility is fixed.

Table A1 summarizes the frozen scoring configuration.

Table A1: **Frozen candidate-scoring configuration used for both datasets.**

Component	Frozen setting
Parser	Stanza 1.11.0; English tokenize, POS, lemma, and dependency processors
Lexical evidence	NRC-VAD v2.1; exact-match gate $\lvert v\rvert \geq 0.5$; intensity $(a + 1)/2$
Context model	Qwen2.5-14B FP16; temperature 0; fixed output schema
Semantic retrieval	BGE-small-en-v1.5; cosine-similarity threshold 0.72

Overlap resolution is applied after eligibility and scoring. When eligible expressions overlap, the longer licensed expression is retained and contained subspans are suppressed. The resulting non-overlapping evidence units remain fixed across severity levels, and their contextual affective intensity scores determine deletion priority in descending order.

Ground-truth sentiment labels and downstream-model predictions are not supplied to candidate extraction, contextual scoring, or deletion-priority assignment. They are used only for downstream evaluation after the corruption artifacts have been frozen.

A.3 SEVERITY DEFINITION

For split s and severity $q \in \{0.2, 0.4, 0.6\}$, we define the *budget specification*

$$\mathbf{b}_{s,q} = \left[C_s, K_{s,q}\right] = \left[\sum_{E \in \mathcal{E}_s} |E|, \operatorname{round}\left(q \sum_{E \in \mathcal{E}_s} |E|\right)\right], \tag{9}$$

where s indexes a dataset split, $\mathcal{E}_s$ is the collection of retained evidence units, C_s is the eligible-word capacity, and $K_{s,q}$ is the target number of deleted surface words.

The three severity values correspond to AER20, AER40, and AER60. Evidence units are selected in descending contextual affective intensity under the split-level budget, with deterministic tie breaking. The realized deletion counts match the target budgets, and each higher-severity deletion set extends the lower-severity set.

A.4 Matched Random Controls

Budget-Matched Random. Budget-Matched Random applies a seeded random ordering to the split-level universe of non-punctuation surface-word positions. Each severity deletes a nested prefix with length equal to the SEL-derived absolute budget $K_{s,q}$. Thus, SEL and Budget-Matched Random have identical split-level deleted-word counts.

Structure-Matched Random. For each SEL-affected utterance admitting a valid alternative, Structure-Matched Random samples spans with the same total deleted-word count, number of spans, and span-length vector. Alternative spans are mutually non-overlapping and disjoint from the corresponding SEL spans. Common-support utterances are fixed independently of corruption draws, and three valid random alternatives are generated for each condition. All matched comparisons evaluate SEL and the random controls on identical common-support utterance IDs.

A.5 Human Audit Protocol and Quality Control

The audit sample contains 60 SEL–Structure-Matched Random pairs from each dataset–severity cell, yielding 360 pairs before quality control. Each pair holds the utterance, deleted-word count, number of spans, and span-length vector fixed. Three independent English-proficient annotators evaluate every pair on a five-point scale indicating which deletion makes sentiment harder to determine, with the midpoint representing a tie. A/B order is randomized independently across annotators and items. Annotators do not observe the condition identity, sentiment label, model prediction, or downstream model.

Post-collection quality control identified a display-layer issue affecting 45 multi-span pairs. The underlying deletion indices remained correct, but the displayed span markers were misaligned. All judgments for these pairs were therefore excluded using a deterministic criterion independent of annotation outcomes. Additional prespecified invalid-item filtering leaves 315 pairs and 932 valid judgments.

Responses are converted to an oriented ordinal score in $\{-2, -1, 0, 1, 2\}$, where positive values favor SEL. Confidence intervals use 20,000 item-clustered bootstrap replicates. Table A2 shows that every dataset–severity interval lies above zero. Overall, 82.4% of valid judgments favor SEL, 11.9% indicate a tie, and 5.7% favor Structure-Matched Random. Ordinal Krippendorff's α is 0.278, indicating stronger agreement on the aggregate direction than on the precise ordinal strength.

Table A2: **Blinded human validation of the SEL relevance proxy.** Positive scores indicate that the SEL deletion makes sentiment harder to determine than its Structure-Matched Random control. Confidence intervals use item-clustered bootstrap resampling. SEL/Tie/Random are percentages of valid judgments.

Dataset	AER	Pairs	Valid	Mean	95% CI	SEL/Tie/Random (%)
MOSI	20	56	167	1.551	[1.353, 1.731]	85.6/7.2/7.2
MOSI	40	56	167	1.569	[1.399, 1.723]	86.2/10.2/3.6
MOSI	60	52	155	1.535	[1.335, 1.714]	86.5/8.4/5.2
MOSEI	20	52	153	1.275	[1.078, 1.458]	75.2/20.3/4.6
MOSEI	40	50	147	1.238	[1.021, 1.435]	76.9/14.3/8.8
MOSEI	60	49	143	1.392	[1.200, 1.569]	83.2/11.9/4.9
Overall	–	315	932	1.432	[1.352, 1.509]	82.4/11.9/5.7

A.6 Benchmark Statistics and Artifact Audit

Table A3 summarizes the test-set statistics for SEL and the two matched random controls. Across all dataset–severity conditions, Structure-Matched Random retains 97.99%–99.66% of SEL-affected utterances in the common support, indicating that the stricter structural control excludes only a small fraction of affected samples.

All frozen corruption artifacts were checked for reconstruction consistency before evaluation, including deletion counts, nested severity structure, and matched-support identities.

Table A3: **Test-set benchmark statistics.** Delete is the full-text deletion rate, Random aff. is the number of utterances affected by Budget-Matched Random, Common is the structure-matched common-support size, and Coverage is its percentage of SEL-affected utterances.

Dataset	AER	Delete (%)	SEL aff.	Random aff.	Common	Coverage (%)
MOSI	20	1.37	92	115	91	98.91
	40	2.73	181	208	179	98.90
	60	4.11	249	277	244	97.99
MOSEI	20	0.90	590	776	588	99.66
	40	1.80	1107	1382	1103	99.64
	60	2.70	1505	1890	1498	99.53

A.7 Complete Robustness Results

Tables A4 and A5 provide the complete model-level Non0-Acc2 results for the two controlled comparisons. The budget-matched comparison uses the full test split, while the structure-matched comparison uses condition-specific common support. For Structure-Matched Random, metrics are averaged across the three corruption draws within each training seed and then across the three training seeds.

Table A4: **SEL versus Budget-Matched Random on the full test split.** Values are three-seed mean Non0-Acc2 (%). Parentheses give SEL−Random in percentage points.

Model	Condition	MOSI			MOSEI		
		AER20	AER40	AER60	AER20	AER40	AER60
Self-MM	Budget-Random	84.1	84.0	83.2	85.1	83.8	84.9
	SEL	**81.5 (-2.7)**	**80.9 (-3.1)**	**78.5 (-4.7)**	**84.1 (-1.0)**	**82.2 (-1.6)**	**82.8 (-2.1)**
MMIM	Budget-Random	83.6	84.0	83.4	82.3	81.4	82.0
	SEL	**80.5 (-3.0)**	**81.1 (-2.9)**	**77.9 (-5.5)**	**81.2 (-1.0)**	**79.9 (-1.5)**	**79.9 (-2.1)**
MISA	Budget-Random	82.7	81.0	82.3	84.8	84.0	84.7
	SEL	**79.9 (-2.7)**	**78.3 (-2.7)**	**78.9 (-3.4)**	**83.9 (-0.8)**	**82.4 (-1.7)**	**82.4 (-2.3)**
ALMT	Budget-Random	82.8	83.1	82.5	85.4	85.4	85.2
	SEL	**79.6 (-3.3)**	**79.1 (-4.0)**	**77.4 (-5.1)**	**84.7 (-0.8)**	**83.8 (-1.6)**	**83.1 (-2.1)**
CENET	Budget-Random	83.2	82.8	83.6	85.0	84.7	85.1
	SEL	**79.9 (-3.3)**	**78.9 (-3.9)**	**78.0 (-5.6)**	**84.0 (-1.0)**	**83.2 (-1.5)**	**82.8 (-2.4)**
MulT	Budget-Random	79.3	79.7	79.8	83.4	83.3	83.4
	SEL	**77.6 (-1.6)**	**77.1 (-2.5)**	**75.9 (-4.0)**	**82.7 (-0.7)**	**81.7 (-1.7)**	**81.2 (-2.3)**
LNLN	Budget-Random	82.6	82.4	82.6	85.1	84.9	85.0
	SEL	**79.3 (-3.2)**	**78.7 (-3.7)**	**77.0 (-5.5)**	**84.4 (-0.7)**	**83.3 (-1.6)**	**82.8 (-2.2)**
TF-Mamba	Budget-Random	82.4	82.8	82.5	84.9	84.6	84.3
	SEL	**79.2 (-3.2)**	**79.1 (-3.7)**	**77.4 (-5.1)**	**83.9 (-0.9)**	**82.6 (-2.0)**	**82.2 (-2.1)**
RECAP-MSA	Budget-Random	80.9	72.8	81.9	85.4	84.7	84.8
	SEL	**77.4 (-3.5)**	**70.2 (-2.6)**	**77.4 (-4.5)**	**84.6 (-0.8)**	**83.2 (-1.5)**	**82.5 (-2.3)**
EBMC	Budget-Random	74.5	75.4	75.8	80.0	79.2	80.0
	SEL	**72.6 (-1.9)**	**73.3 (-2.0)**	**72.0 (-3.8)**	**78.9 (-1.1)**	**77.6 (-1.5)**	**77.8 (-2.2)**
Mean	Budget-Random	81.6	80.8	81.7	84.1	83.6	83.9
	SEL	**78.8 (-2.8)**	**77.7 (-3.1)**	**77.0 (-4.7)**	**83.3 (-0.9)**	**82.0 (-1.6)**	**81.7 (-2.2)**

A.8 Mechanism-Matched Training

Figure A1 covers all 60 model–dataset–severity conditions. Both training conditions are evaluated on the same full SEL test split.

Table A5: **SEL versus Structure-Matched Random on condition-specific common support.** Values are mean Non0-Acc2 (%). Parentheses give SEL−Random in percentage points.

Model	Condition	MOSI			MOSEI		
		AER20	AER40	AER60	AER20	AER40	AER60
Self-MM	Structure-Random	86.0	86.5	85.6	86.8	87.2	87.0
	SEL	**64.8 (-21.1)**	**71.7 (-14.8)**	**70.0 (-15.6)**	**79.6 (-7.2)**	**80.8 (-6.4)**	**80.9 (-6.1)**
MMIM	Structure-Random	86.1	86.4	84.6	86.1	85.6	84.4
	SEL	**62.3 (-23.8)**	**74.1 (-12.3)**	**68.5 (-16.1)**	**77.8 (-8.4)**	**78.5 (-7.0)**	**77.8 (-6.5)**
MISA	Structure-Random	85.7	81.7	83.5	87.5	86.5	87.6
	SEL	**61.9 (-23.8)**	**69.5 (-12.2)**	**72.0 (-11.5)**	**79.6 (-8.0)**	**79.1 (-7.5)**	**80.5 (-7.1)**
ALMT	Structure-Random	86.6	85.0	83.9	87.4	87.5	87.3
	SEL	**60.4 (-26.1)**	**69.5 (-15.5)**	**69.8 (-14.1)**	**80.6 (-6.9)**	**79.9 (-7.5)**	**80.9 (-6.5)**
CENET	Structure-Random	84.6	84.3	85.5	86.9	87.1	87.5
	SEL	**59.3 (-25.3)**	**68.5 (-15.8)**	**69.6 (-15.9)**	**79.2 (-7.7)**	**80.7 (-6.5)**	**80.7 (-6.8)**
MulT	Structure-Random	78.3	79.0	78.6	85.6	85.5	85.3
	SEL	**66.7 (-11.6)**	**68.0 (-11.0)**	**67.6 (-10.9)**	**79.9 (-5.6)**	**78.9 (-6.7)**	**79.3 (-6.0)**
LNLN	Structure-Random	87.2	84.5	83.9	87.6	87.5	87.7
	SEL	**62.3 (-24.9)**	**69.5 (-15.1)**	**68.7 (-15.2)**	**80.2 (-7.4)**	**80.5 (-7.0)**	**80.4 (-7.3)**
TF-Mamba	Structure-Random	87.5	84.4	84.3	88.3	87.7	86.9
	SEL	**62.3 (-25.3)**	**69.3 (-15.1)**	**69.3 (-15.0)**	**80.4 (-7.9)**	**79.9 (-7.9)**	**81.0 (-5.9)**
RECAP-MSA	Structure-Random	86.9	75.8	84.5	87.4	87.4	87.1
	SEL	**60.8 (-26.1)**	**65.2 (-10.6)**	**69.8 (-14.6)**	**80.2 (-7.2)**	**80.8 (-6.6)**	**80.5 (-6.5)**
EBMC	Structure-Random	78.6	75.7	76.3	83.4	82.9	82.7
	SEL	**60.4 (-18.2)**	**66.1 (-9.6)**	**63.6 (-12.6)**	**75.3 (-8.1)**	**76.3 (-6.6)**	**76.6 (-6.1)**
Mean	Structure-Random	84.7	82.3	83.1	86.7	86.5	86.3
	SEL	**62.1 (-22.6)**	**69.1 (-13.2)**	**68.9 (-14.2)**	**79.3 (-7.4)**	**79.5 (-7.0)**	**79.9 (-6.5)**

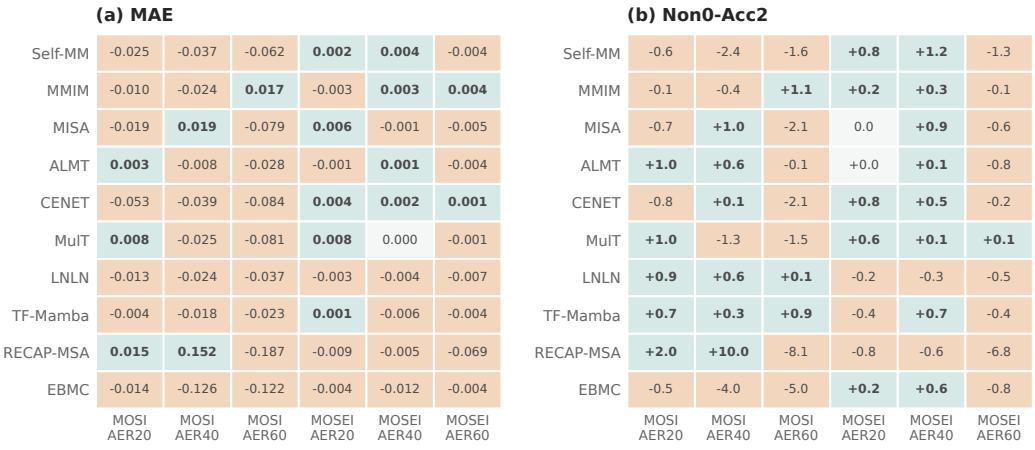


Figure A1: **Condition-level effects of mechanism-matched training on the full SEL test split.** Panel a presents Budget-Random-trained minus SEL-trained MAE, and panel b shows SEL-trained minus Budget-Random-trained Non0-Acc2. Positive values favor SEL-matched training. Improvements occur in 17 of 60 conditions for MAE and 29 of 60 for Non0-Acc2.

B SRDP IMPLEMENTATION AND REPRODUCIBILITY

B.1 SRDP CONFIGURATION

We fix $\lambda_{\text{corr}} = \lambda_{\text{dist}} = \lambda_{\text{pred}} = \lambda_{\text{feat}} = 0.1$ across all evaluated SRDP configurations. A subset-specific auxiliary objective is set to zero when its required subset is empty, and $\mathcal{L}_{\text{rel}} = 0$ when fewer than two affected samples occur in a mini-batch. The native objective, trainable parameter set, optimizer, and model-specific training stages otherwise follow the paired Native configuration.

For an affected pair (i, j), define the prediction residual of sample k with respect to its clean-view reference as

$$\varepsilon_k = p^m_{S,k} - p^c_{R,k}, \qquad \delta^m_{ij} - \delta^c_{ij} = \varepsilon_i - \varepsilon_j. \tag{10}$$

The prediction-relation term therefore constrains differences among affected-sample residuals, while the native task objective anchors absolute predictions.

B.2 TRAINING SETUP AND CHECKPOINT SELECTION

We use the unaligned MOSI and MOSEI splits, containing 1,284/229/686 and 16,326/1,871/4,659 train/validation/test utterances, respectively. The evaluated models include six conventional multimodal sentiment models: Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), MISA (Hazarika et al., 2020), ALMT (Zhang et al., 2023), CENET (Wang et al., 2023), and MulT (Tsai et al., 2019); and four models designed for incomplete multimodal inputs: LNLN (Zhang et al., 2024), TF-Mamba (Li et al., 2025), RECAP-MSA (Huang et al., 2026), and EBMC (He et al., 2026b). All formal results use training seeds 1111, 1112, and 1113.

Table A6 summarizes the backbone-specific settings used for the ten-model MOSEI AER60 comparison. Within each model–seed pair, Native and SRDP use the same optimizer, batch size, epoch control, validation-only checkpoint rule, and deterministic initialization. SRDP adds only the objectives defined in Section 4.

Table A6: **Training configuration for the ten-model MOSEI AER60 comparison.** Learning rates are listed by parameter group when a model uses more than one. "Update" denotes gradient-accumulation steps, and "LR patience" denotes learning-rate scheduler patience.

Model	Optimizer and learning rate	Batch	Epoch control	Validation selection
Self-MM	Adam; BERT 5e−5, A 5e−3, V 1e−4, other 1e−3	32	early stop 8; update 4	loss
MMIM	Adam; BERT 5e−5, main/MMILB 1e−3	32	early stop 8; update 2	loss
MISA	Adam; 1e−4, weight decay 5e−5	64	early stop 8; update 2	MSE
ALMT	AdamW; 1e−4, weight decay 1e−4	64	max 200; early stop 32	MSE
CENET	Adam; 1e−5, weight decay 1e−4	64	early stop 8	L1
MulT	Adam; 5e−4, weight decay 1e−3	4	early stop 8; LR patience 20; update 8	loss
LNLN	AdamW; 1e−4, weight decay 1e−4	64	200 epochs	minimum MAE, rounded to 4 d.p.
TF-Mamba	AdamW; 1e−4, weight decay 1e−4	64	200 epochs	minimum MAE, rounded to 4 d.p.
RECAP-MSA	AdamW; 1e−4, weight decay 1e−4	64	200 epochs	minimum unrounded MAE
EBMC	Adam; 1e−4, weight decay 1e−5	16	100 epochs; stage II from 50	F1

For the SEL robustness experiments, only Budget-Matched Random-trained and SEL-trained checkpoints are trained. Structure-Matched Random is used only as a test-time control with the corresponding frozen Budget-Matched Random-trained checkpoint.

Each model retains its native validation-only checkpoint-selection rule. The MMSA implementations use their original loss-based criteria, including L1 for CENET and MSE for ALMT and MISA, while EBMC selects checkpoints by validation F1. Test data are not used for epoch selection, hyperparameter tuning, method selection, or selection of the midpoint coefficient.

For SRDP, the selected Native checkpoint serves as the frozen preservation anchor and one endpoint of checkpoint integration. The reference model is separately trained and validation-selected for the same backbone, dataset, severity level, and seed. Both the Native anchor and reference model remain frozen during SRDP training.

C Additional SRDP Analyses

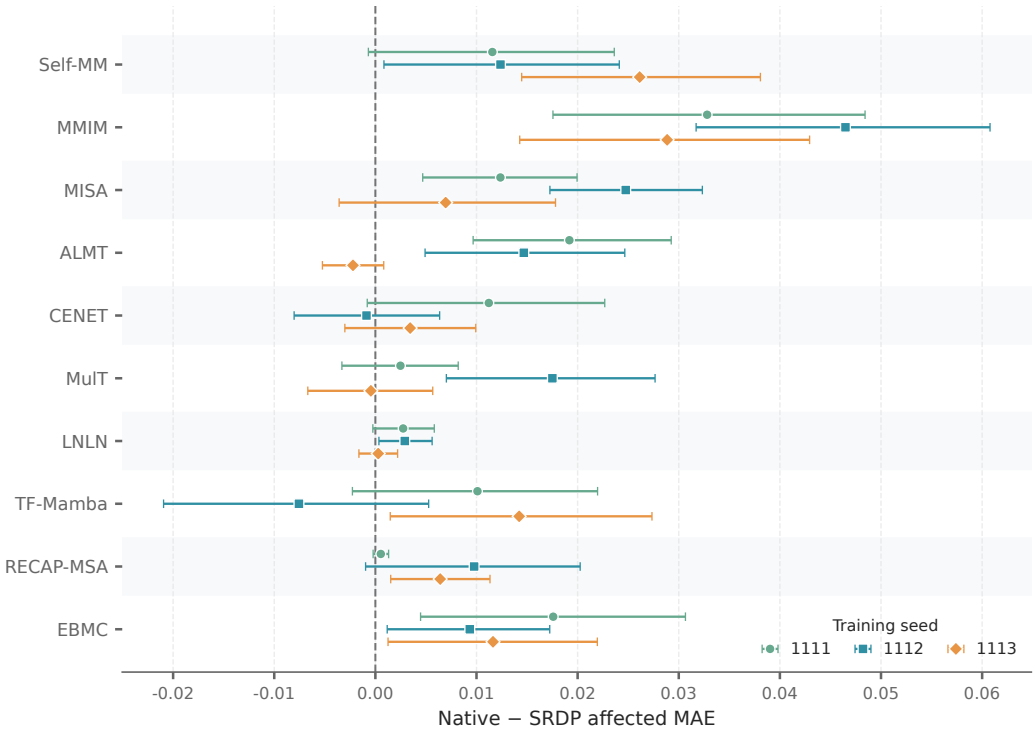


Figure A2: **Seed-specific uncertainty for Affected MAE on MOSEI at AER60.** Points show Native minus SRDP for each model and training seed, so positive values favor SRDP. Bars are 95% intervals from 10,000 paired video-cluster bootstrap replicates within the corresponding fixed seed.

C.1 Training Controls and Component Ablations

Table A7 provides the complete Self-MM training-strategy controls on MOSEI at AER60. Differences are computed within each training seed before aggregation.

Table A7: **Training-strategy controls on Self-MM at MOSEI AER60.** Values are mean ± sample standard deviation over three paired seeds. Differences are method−Native. MAE differences are absolute, and Non0-Acc2 differences are percentage points.

Training condition	Δ Full MAE	Δ Affected MAE	Δ Unaffected MAE	Δ Full Non0-Acc2
Clean+Corrupted Joint	-0.0137 ± 0.0045	-0.0039 ± 0.0098	-0.0183 ± 0.0025	$+1.49 \pm 1.99$
Pointwise distillation	-0.0148 ± 0.0014	-0.0130 ± 0.0024	-0.0157 ± 0.0024	$+1.42 \pm 1.44$
SRDP	-0.0146 ± 0.0022	-0.0167 ± 0.0082	-0.0136 ± 0.0051	$+0.28 \pm 1.61$
Corrupted-view reference	-0.0183 ± 0.0005	-0.0168 ± 0.0037	-0.0190 ± 0.0022	$+1.09 \pm 1.72$

SRDP produces a larger Affected MAE reduction than clean–corrupted joint training and pointwise distillation in this comparison. The validation component ablations summarized in Table 2 further show that shuffling pairwise correspondence or removing Native preservation increases Affected MAE. Removing representation relations produces a smaller degradation.

We additionally examine whether SRDP's improvement depends specifically on feeding the clean view to the frozen reference. For this control, only the reference input view is changed while the remaining training and checkpoint protocol is held fixed. Table A8 presents the added MMIM and TF-Mamba comparisons; the corresponding Self-MM result appears in Table A7.

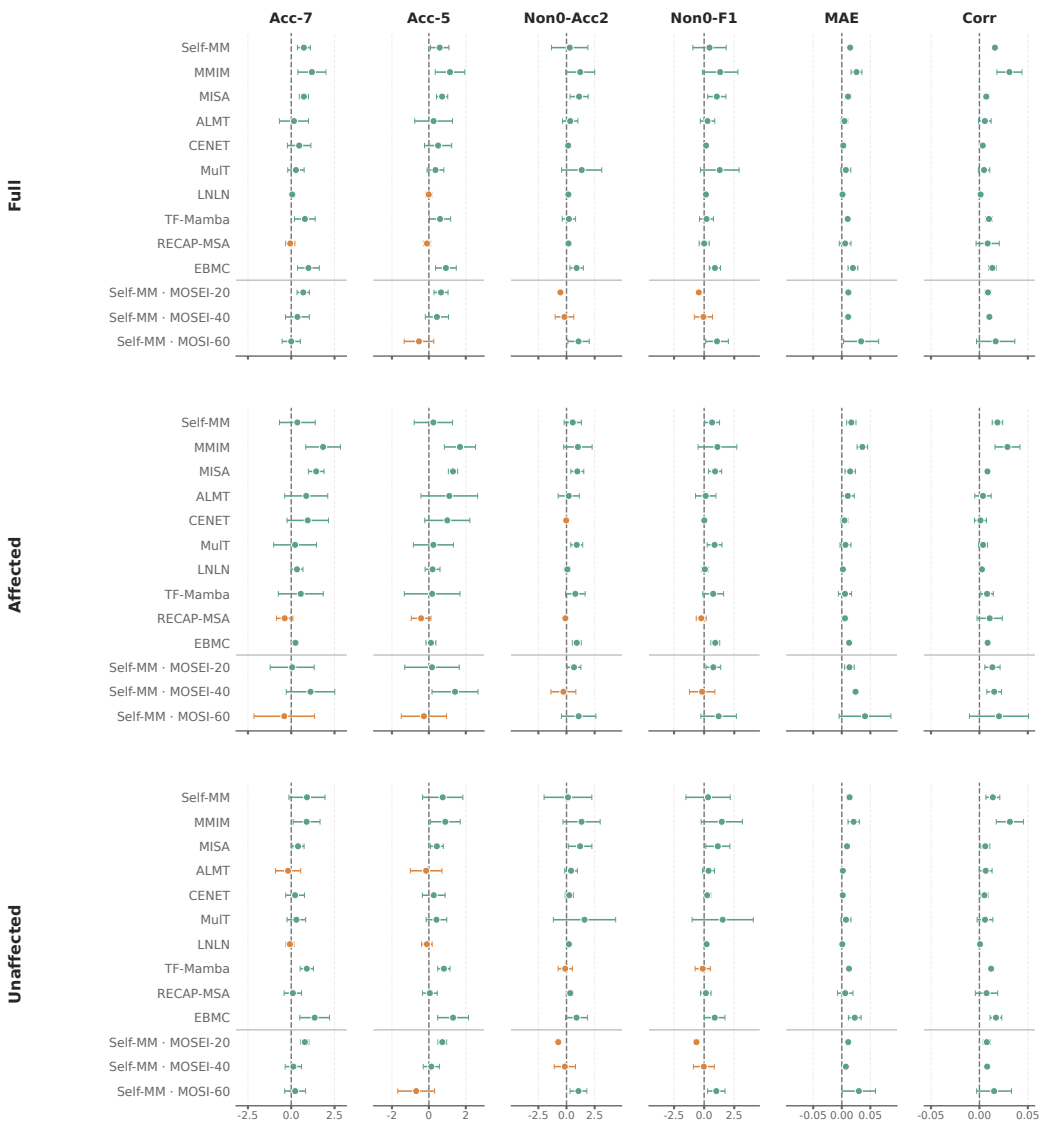


Figure A3: **Native-to-SRDP changes across metrics and evaluation subsets.** Each point is one model–dataset–severity mean over three paired seeds, and whiskers show the sample standard deviation of the paired differences. Acc-7, Acc-5, Non0-Acc2, and Non0-F1 use SRDP−Native in percentage points; Corr uses SRDP−Native; MAE uses Native−SRDP. Rightward values therefore favor SRDP throughout.

Table A8: **Clean- versus corrupted-reference control on MOSEI at AER60.** Values are three-seed mean Affected MAE for the fixed midpoint. The difference is Clean−Corrupted, and lower MAE is better.

Backbone	Clean ref.	Corrupted ref.	Difference	Seed direction	Stable clean advantage
MMIM	0.7739	0.7820	−0.0081	Clean lower, 3/3	No
TF-Mamba	0.7585	0.7566	+0.0019	Mixed	No

A stable clean-reference advantage requires a consistent seed direction with seed-specific uncertainty excluding zero. MMIM has the same directional advantage for the clean reference in all three seeds, but the uncertainty criterion is not met consistently; TF-Mamba has mixed seed directions.

Together with the Self-MM control, these results do not establish the clean reference view as the source of SRDP's improvement. They do not imply equivalence between the two reference views.

C.2 Checkpoint Integration and Uncertainty

The deployment checkpoint is constructed after Native and SRDP checkpoint selection using the fixed interpolation coefficient of 0.5. The coefficient is not selected using validation or test performance. The main-text checkpoint analysis shows that the fixed midpoint lowers mean MAE relative to Native for all ten models on MOSEI AER60 and generally improves upon the selected SRDP student across the Full, Affected, and Unaffected subsets.

Figure A2 further shows seed-specific uncertainty for Affected MAE across all ten models. Each interval resamples test videos within a fixed training seed, separating test-sample uncertainty from variation across training seeds.

C.3 Complete Metric Profiles

Figure A3 presents Native-to-SRDP changes across all reported metrics for the Full, Affected, and Unaffected subsets. The 13 model–dataset–severity conditions are retained in each panel, allowing the primary Affected MAE result to be viewed together with secondary metrics and subset-specific trade-offs.