BEYOND SUMMARIES: LEARNING TO READ LOCAL EVIDENCE FOR INCOMPLETE MULTIMODAL SENTIMENT ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Incomplete multimodal sentiment analysis predicts sentiment from partially observed language, audio and visual inputs. Existing approaches for incomplete multimodal sentiment analysis predominantly rely on global representations, yet overlook the heterogeneous performance degradation induced by the loss of distinct local regions. Even under identical missing ratios, the absence of decisive local cues may completely flip the predicted sentiment polarity. To mitigate such drawbacks, this paper presents DREAM, a damage-guided framework for discovering cross-modal local evidence. Through contrastive analysis of prediction risks triggered by paired regional erasures, DREAM quantifies the task-level relevance of multi-scale local evidence while preserving global summary information. Equipped with a summary-referenced reader module, each modality adaptively balances a source modality's global summary with its complementary local evidence, conditioned on the receiving modality's available content. Extensive evaluations on MOSI, MOSEI and CH-SIMS demonstrate consistent improvements in both sentiment classification and regression. Performance gains are especially substantial on CH-SIMS, where missing fragments exert highly uneven impacts on sentiment polarity. Without full-view teachers or supervision from fully observed counterparts, DREAM improves binary classification accuracy and F1 over MIDAS by 4.11 and 5.05 percentage points, respectively, while reducing MAE by 12.72%.

1 INTRODUCTION

Multimodal sentiment analysis combines complementary language, acoustic, and visual cues to infer sentiment. With partially missing observations, incomplete multimodal sentiment analysis must rely on the remaining evidence (Wu et al., 2026). Recent work addresses unequal modality availability during training (Nguyen et al., 2026) and learning across different modality combinations (Zhao et al., 2025; Xu et al., 2026). Beyond which modalities are available, robustness depends on which content is lost. At the same missing rate, losing a decisive phrase or brief affective cue can reverse predicted polarity, whereas removing less informative fragments may leave it unchanged. These unequal consequences make local task value central to effective use of incomplete observations.

Existing approaches largely organize this information through restored features and modality-level representations. Feature completion and complete-view supervision strengthen incomplete representations (He et al., 2026; Zou et al., 2026), while MIDAS (Wen et al., 2026) disentangles shared and modality-specific information for uncertainty-aware fusion, and P-RMF (Zhu et al., 2025) combines probabilistic proxies with dynamic cross-modal injection. These approaches improve representation quality, but modality-level uncertainty does not directly specify the consequences of losing a particular region. Even when broad context remains recognizable, the missing fragment may determine sentiment polarity. This motivates learning local evidence through its task consequences, beyond relying on global summaries or feature compatibility alone.

The importance of local evidence is reflected in prediction failures associated with missing key frames and underused affective cues (Bao et al., 2026; Sun et al., 2026). Figure 1 makes the local distinction concrete: two masks delete the same number of tokens from a clip, yet retaining or removing "exciting bit" produces opposite predicted polarities. Equal deletion budgets therefore need not imply equal

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.

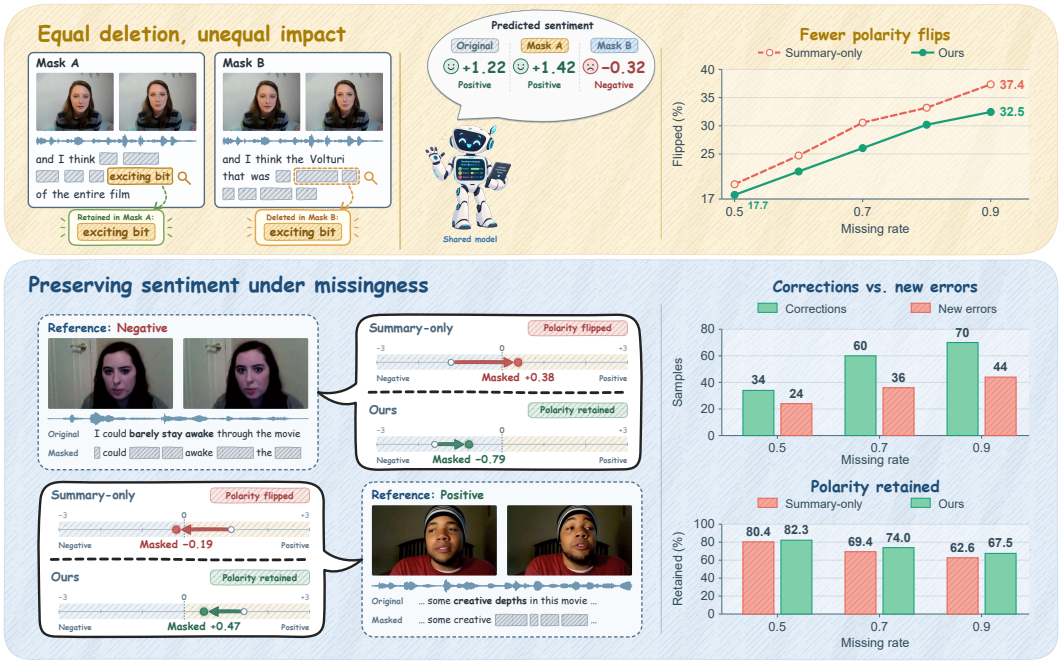

Figure 1: Equal deletion, unequal impact on sentiment. Top left: equal-budget text deletions yield opposite polarities in a recorded MOSI training example. Bottom left: DREAM preserves polarity in selected incomplete test examples. Right: DREAM yields fewer flips than Summary-only on a shared 530-sample test cohort at 50%–90% missingness.

task damage. Our insight is to turn this unequal damage into supervision: learn the value of a region from how its removal changes prediction risk. Because decisive cues can occupy a short expression or a broader contextual span, this supervision should preserve access to evidence at multiple scales.

Building on this insight, we propose DREAM (**D**amage-guided **R**eading of **E**vidence **A**cross **M**odalities) for incomplete multimodal sentiment analysis, connecting learned local task value with cross-modal evidence reading. Our contributions are:

- We propose **damage-guided multiscale evidence learning** to distinguish sentiment-critical regions whose unequal impact is obscured by focusing only on the amount of missing information. Paired, equal-budget erasures provide task-sensitive supervision for learning regional value across scales.
- We develop a **summary-referenced cross-modal reader** to address the loss of local cues in summary-only exchange. It combines learned regional value with cross-modal compatibility, enabling each modality to select complementary local evidence from the others while preserving access to global summaries.
- We demonstrate **robustness gains from connecting evidence value with evidence use** on MOSI, MOSEI, and CH-SIMS under random and structured missingness. On CH-SIMS, DREAM improves binary accuracy and F1 over MIDAS by 4.11 and 5.05 percentage points, respectively, while reducing MAE by 12.72%. Multiple component ablations link the gains to evidence learning and reading, supported by full-test comparisons and qualitative examples.

2 Related Work

2.1 Incomplete Multimodal Sentiment Analysis

Multimodal sentiment representations must retain both shared and distinctive signals. Shared–specific factorization in MISA (Hazarika et al., 2020) and modality-specific supervision in Self-MM (Yu et al.,

2021) pursue this balance, while MMIM (Han et al., 2021) preserves information through hierarchical mutual-information maximization. Missing observations make retaining such information harder, motivating explicit recovery and robust representation learning.

Recovery supplies representations that incomplete inputs cannot directly provide. TFR-Net (Yuan et al., 2021) and EMT-DLFR (Sun et al., 2024) combine contextual interaction with feature reconstruction. Recent designs refine what is reconstructed: RECAP (Huang et al., 2026) separates multigranular temporal reconstruction from shared–specific semantic factorization, while SMCIR (Chen et al., 2026) couples sample-level missingness diagnosis with text-guided completion. Disentangled recovery (Yang et al., 2025) and token-level information bottlenecks with cyclic translation (Lin et al., 2025b) further organize task-relevant content. These objectives emphasize recovering or restructuring representations, leaving their local task value as a distinct learning target.

The available training reference also shapes this target. Complete-view information supports teacher distillation (Luo et al., 2025; Zhuang et al., 2025), predictive anchoring (He et al., 2026), and boundary-aware relational transfer (Zhang et al., 2025); denoising can also precede compensation through a complete-data training stage (Zhuang et al., 2026). In CMAD, teacher–student task-loss differences determine modality-combination weights (Zhuang et al., 2025). Such references guide incomplete representations at the view or combination level. DREAM instead derives regional supervision directly from matched additional erasures within an incomplete observation, making the task consequences of local information loss explicit.

Representation quality must also translate into useful fusion. LNLN (Zhang et al., 2024) compensates for degraded language, while MIDAS (Wen et al., 2026) and P-RMF (Zhu et al., 2025) use uncertainty in disentangled representations and probabilistic proxies. Reliability estimation additionally draws on reconstruction error (Zou et al., 2026) and variance-modulated experts (Zhu et al., 2026). Observed-only reconstruction regularizes experts (Hussain et al., 2026), while sentiment-intensity supervision and uncertainty regulate modality contributions (Cao et al., 2026). These strategies address useful information at the modality level. A modality can nevertheless remain broadly reliable while losing the particular cue another modality needs. Fusion thus benefits from a finer-grained account of what each source can contribute.

2.2 Local Evidence and Cross-Modal Interaction

Reading individual regions requires access beyond modality summaries. CENET (Wang et al., 2023b) enriches language with nonverbal context, while TETFN (Wang et al., 2023a) uses text to enhance nonverbal representations and ALMT (Zhang et al., 2023) filters nonverbal information through multiscale language guidance. More recent approaches combine global and local fusion (Tao et al., 2025), align token-level and distribution-level information (Li et al., 2025b), or query audio-visual segments with text (Li et al., 2025a). DNRNet couples local–global recurrent queries with missing-rate-based confidence supervision (Zhang et al., 2026). These mechanisms establish access to local information; task-damage supervision provides a basis for deciding which of that information matters to sentiment.

That decision also depends on the supervision attached to evidence. Attention rankings (Mai et al., 2025), unimodal prediction-loss-derived quality (Lin et al., 2025a), and masked teacher outputs (Covert et al., 2025) supervise different aspects of information selection or local representation. PRLF examines key-frame missingness while estimating modality importance through confidence and Fisher information (Bao et al., 2026). DREAM focuses on a region's incremental task damage in its existing incomplete context. Matching deletion budgets makes the supervision sensitive to what is removed, connecting regional value to the prediction consequences that motivate our design.

Learned value must then inform context-dependent use. Retrieval methods consult external instances to compensate for missing modalities (Pipoli et al., 2025; Lang et al., 2025), while adaptive selection can acquire new features (Guney et al., 2025) or choose recovered modalities (Du et al., 2026). DREAM reads evidence already available within the same example. Its summary-referenced reader combines source-side task value with cross-modal compatibility, selecting between each source's summary and multiscale local evidence. The summary provides a contextual reference, allowing each modality to favor local evidence when its current content calls for more specific information. This connection between learning evidence value and deciding how to read it defines the method developed next.

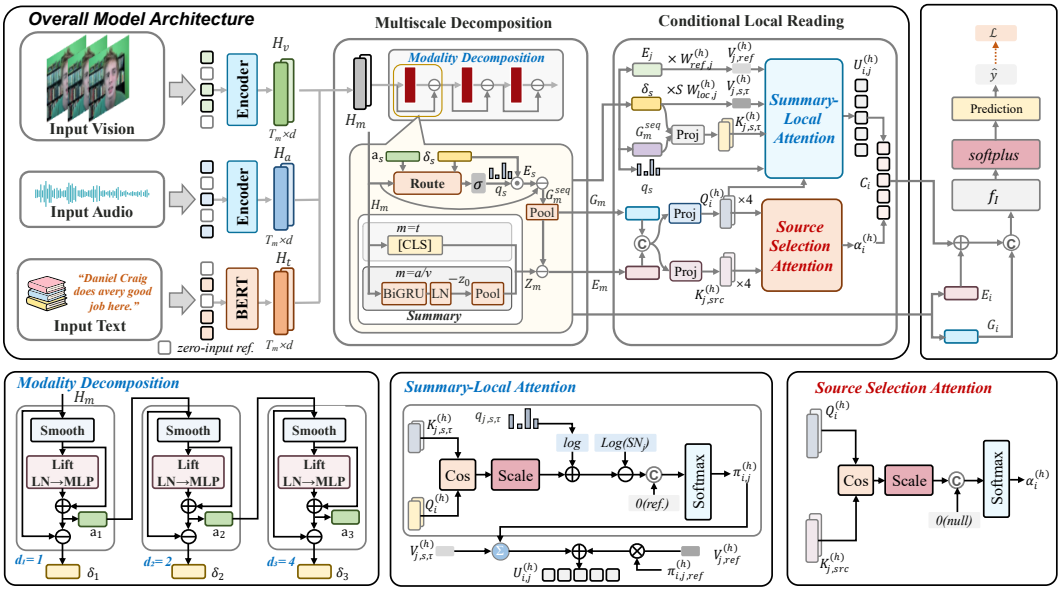

Figure 2: DREAM overview. Branch labels give dilations. Prediction combines intensity and signed polarity (Section 3.4); matched-erasure training is detailed in Section 3.2.

3 METHOD

For text, audio, and vision, indexed by $m \in \mathcal{M} = \{t, a, v\}$, DREAM maps incomplete sequences $\widetilde{x} = \{\widetilde{x}_m\}$ to sentiment $\widehat{y}$ with target y. Encoders yield $H_m \in \mathbb{R}^{T_m \times d}$, with sequence length T_m and shared width d, using incomplete inputs and original padding validity. DREAM learns multiscale evidence value from matched erasures and reads it according to the receiving modality's needs (Figure 2).

3.1 TASK-VALUED MULTISCALE EVIDENCE

Sentiment cues span brief expressions and broader context. From $a_{m,0} = H_m$, stages $s = 1, \ldots, S$ form $a_{m,s}$ by padding-normalized smoothing and learned lifting (Appendix D.3). The residual $\delta_{m,s} = a_{m,s-1} - a_{m,s}$ captures detail at scale s. A learned score $q_{m,s}$, computed using the sigmoid function σ from each position's original, lifted, and residual features, determines how much local detail is allocated away from the global stream:

$$G_m^{\text{seq}} = H_m - \sum_{s=1}^{S} q_{m,s} \odot \delta_{m,s}. \tag{1}$$

Here $\odot$ multiplies position-wise scores across feature channels; subtraction preserves exact recomposition.

The modality-specific summarizer $\mathcal{A}_m$ (implementation in Appendix D.3) maps a sequence to a single vector. Applying the same summarizer gives the global summary $G_m = \mathcal{A}_m(G_m^{\text{seq}})$ and the input summary $Z_m = \mathcal{A}_m(H_m)$, with summary residual $E_m = Z_m - G_m$. Memory retains every valid $\delta_{m,s}$ and $q_{m,s}$, providing access beyond E_m.

3.2 LEARNING VALUE FROM MATCHED ERASURES

Local value should reflect prediction consequences. We erase regions A and B from the same incomplete modality, matching valid width and newly erased count but changing locations. Let $\widetilde{x}^{(-U)}$ denote the input after erasing region U, with other modalities fixed. Deletion damage measures the risk increase:

$$D_U = \text{sg}\Big[\mathcal{R}(\widetilde{x}^{(-U)}, y) - \mathcal{R}(\widetilde{x}, y)\Big], \qquad U \in \{A, B\}. \tag{2}$$

Here $\mathcal{R}(x, y) = (\widehat{y}(x) - y)^2 + \ell_{\text{pol}}(\ell(x), y)$ combines sentiment error with weighted binary cross-entropy on the polarity logit ℓ; sg stops gradients. Positive D_U indicates harmful deletion.

A region can matter through a few decisive positions. Noisy-OR aggregates scores over its newly erased positions, and the resulting logit ρ_U predicts regional value. The pair (A, B) compares the damage of the two regional deletions, while $(A, 0)$ and $(B, 0)$ compare each deletion against no additional erasure. Here 0 denotes the original incomplete input, with reference values $D_0 = \rho_0 = 0$. For sample b, let $\mathcal{P}_b$ contain its eligible $(A, 0)$, $(B, 0)$, and (A, B) pairs. Writing their damage and score differences as $\Delta D_{b,UV}$ and $\Delta\rho_{b,UV}$, we train the ordering by

$$\mathcal{L}_{\text{rank}} = \sum_b \sum_{(U,V)\in\mathcal{P}_b} \overline{w}_{b,UV}\ \text{softplus}[-\,\text{sgn}(\Delta D_{b,UV})\,\Delta\rho_{b,UV}]\,. \tag{3}$$

The sign function sgn specifies the ordering. Weights $\overline{w}_{b,UV}$ are proportional to $|\Delta D_{b,UV}|$ and normalized over the minibatch. When no matched region B is available, we retain $(A, 0)$; equal-damage comparisons have zero weight.

Ranking learns regional value; damage-weighted assignment connects this value to the decomposition. Let $\eta_H^{b,U}$ and $\eta_G^{b,U}$ denote the mean squared changes in H_m and G_m^{seq} after erasing region U in sample b. We penalize the relative global-stream response:

$$\mathcal{L}_{\text{assign}} = \sum_{b,U} \overline{\omega}_{b,U}\ \frac{\eta_G^{b,U}}{\max(\eta_H^{b,U}, \epsilon)}. \tag{4}$$

Here $\overline{\omega}_{b,U}$ are normalized weights proportional to positive deletion damage over eligible regions, and ϵ is a numerical safeguard. Under exact recomposition, this encourages local detail to carry the valuable responses. Evaluation-mode risks and encoder features are detached, so regional objectives train the split with fixed measurements. Appendix D.3 gives the weight definitions, eligibility rules, and gradient paths.

3.3 Summary-Referenced Cross-Modal Reading

Task value must be interpreted in each modality's context. For modality i and source modality $j \neq i$, head h projects a query $Q_i^{(h)}$ from $[G_i; E_i]$ and keys $K_{j,s,\tau}^{(h)}$ from $[G_{j,\tau}^{\text{seq}}; \delta_{j,s,\tau}]$ at position τ; $[\cdot;\cdot]$ concatenates features. With head width d_h, each local score combines cross-modal compatibility and task value:

$$u_{ij,s,\tau}^{(h)} = \sqrt{d_h}\cos(Q_i^{(h)}, K_{j,s,\tau}^{(h)}) + \log q_{j,s,\tau} - \log(SN_j). \tag{5}$$

Here cos is cosine similarity and N_j counts padding-valid positions except text special tokens. The memory-size correction is $\log(SN_j)$. A softmax over local logits and a zero-logit summary reference gives weights π, yielding

$$U_{ij}^{(h)} = \pi_{ij,\text{ref}}^{(h)} W_{\text{ref},j}^{(h)} E_j + \sum_{s,\tau} \pi_{ij,s,\tau}^{(h)} W_{\text{loc},j}^{(h)} (S\delta_{j,s,\tau}). \tag{6}$$

Learned value maps $W_{\text{ref},j}^{(h)}$ and $W_{\text{loc},j}^{(h)}$ project summaries and raw detail; S compensates for scale count. Outer attention scores sources from $[G_j; E_j]$ with the same query from modality i and a zero-valued null option. Combining weighted messages across heads gives C_i, allowing each modality to select sources or decline exchange. Appendix D.3 details source selection and multi-head aggregation.

3.4 Sentiment Prediction and Training

The readout separates direction from magnitude. Polarity heads read $Z_m + C_m$ and global-summary interactions to predict positive-class probability $P_+ = \sigma(\ell)$. A joint head predicts nonnegative intensity I from $[G; E + C]$, with G, E, C collecting modality summaries and messages. The prediction is $\widehat{y} = I(2P_+ - 1)$. Setting $C = 0$ gives $\widehat{y}_0$, preserving a direct observed-information path. Training minimizes

$$\mathcal{L} = \mathbb{E}\big[\mathcal{R}(\widetilde{x}, y) + (\widehat{y}_0 - y)^2 + \lambda_I(I - |y|)^2\big] + \lambda_{\text{aux}}\mathcal{L}_{\text{aux}}. \tag{7}$$

The expectation averages training examples; λ_I weights intensity error. Here $\mathcal{L}_{\text{aux}} = \mathcal{L}_{\text{rank}} + \mathcal{L}_{\text{assign}} + \mathcal{L}_{\text{uni}}$, with $\mathcal{L}_{\text{uni}}$ averaging sentiment error from individual Z_m summaries. All auxiliary terms share coefficient λ_{aux}. Appendix D.3 gives head definitions and fixed settings. Only multimodal labels supervise DREAM; erasure comparisons are training-only.

4 EXPERIMENTS

We first compare DREAM with existing methods across random missing rates and conduct additional tests under structured missingness, which removes contiguous spans or entire modalities. We then combine controlled ablations, quantitative analyses, and qualitative examples to examine how damage-guided learning identifies task-valued local evidence and how summary-referenced reading uses it to support sentiment prediction from incomplete multimodal observations.

4.1 EXPERIMENTAL SETUP

We evaluate on MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). For consistent comparisons, we follow the data splits, features, and random-missingness protocol adopted in MIDAS (Wen et al., 2026). We evaluate DREAM and its ablations across three seeds (1111, 1112, 1113). Erased text becomes `[UNK]`; erased audio/visual features become zero. Each DREAM run selects one checkpoint at validation rate $r = 0.5$ for all test rates $r \in \{0.0, 0.1, \ldots, 0.9\}$. Training and aggregation details are in Appendix D.

We reproduce baseline models, including RECAP and SMCIR, from their released implementations, retaining their core architectures and adapting dataset interfaces to the shared missing-input protocol. Appendix D.4 details baseline adaptations and training settings.

Metrics cover binary/multiclass accuracy, weighted F1, MAE, and correlation. English Acc-2/F1 superscripts $-/+$ exclude/include neutral targets; MAE is lower-is-better.

4.2 PREDICTIVE PERFORMANCE UNDER MISSINGNESS

DREAM achieves the best reported average on every metric across the three benchmarks (Tables 1–3). The gains cover both sentiment polarity and intensity.

On CH-SIMS, DREAM improves Acc-2 and F1 over MIDAS by 4.11 and 5.05 percentage points, while reducing MAE by 12.72%. The advantage also extends to finer distinctions: Acc-3 rises by 5.28 points and Corr increases from 0.427 to 0.559. Thus the improvement spans binary decisions, graded sentiment categories, and continuous scores.

Table 1: CH-SIMS results averaged over ten missing rates. DREAM: mean ± SD. **Bold**/underline: best/second-best.

Method	Acc-2↑	F1↑	Acc-3↑	Acc-5↑	MAE↓	Corr↑
MISA (MM'20)	64.79	63.20	51.45	30.03	0.533	0.368
Self-MM (AAAI'21)	69.40	68.60	53.18	31.44	<u>0.507</u>	0.391
TFR-Net (MM'21)	63.27	59.22	49.96	26.98	0.593	0.192
MMIM (EMNLP'21)	66.99	64.97	50.95	27.81	0.552	0.306
CENET (TMM'23)	61.76	57.33	48.54	22.44	0.605	0.172
TETFN (PR'23)	68.34	67.78	51.16	31.26	0.510	0.386
ALMT (EMNLP'23)	68.81	68.14	53.02	32.29	0.542	0.326
EMT-DLFR (TAC'24)	71.41	71.09	57.66	<u>34.53</u>	0.517	0.403
LNLN (NeurIPS'24)	69.43	68.04	54.33	31.24	0.540	0.323
P-RMF (ACL'25)	71.31	68.76	53.44	30.56	0.525	0.377
RECAP (AAAI'26)	68.74	67.46	54.70	29.72	0.539	0.287
SMCIR (AAAI'26)	69.16	68.75	54.93	31.36	0.526	0.367
MIDAS (TPAMI'26)	<u>73.82</u>	<u>72.32</u>	<u>58.69</u>	33.12	<u>0.507</u>	<u>0.427</u>
DREAM	**77.94**±1.75	**77.37**±1.39	**63.97**±1.38	**38.70**±0.82	**0.442**±0.008	**0.559**±0.021

On MOSI, DREAM reduces MAE by 5.29% and improves Acc-7 from 30.76% to 33.41% (Table 2).

On the larger MOSEI benchmark, DREAM also improves every average metric, including a 0.63-point gain in F1 with neutral targets included (Table 3).

Retaining performance as observations disappear. Figure 3 shows how these gains evolve with the missing rate. DREAM leads all three displayed metrics on CH-SIMS throughout the range. At 90% missingness, its F1 reaches 70.97%, 58.86%, and 65.07% on CH-SIMS, MOSI, and MOSEI.

Table 2: MOSI results averaged over ten missing rates. DREAM: mean ± SD. Superscripts − and + exclude and include neutral targets.

Method	Acc-2$^-$ ↑	Acc-2$^+$ ↑	F1$^-$ ↑	F1$^+$ ↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
MISA (MM'20)	70.89	68.87	67.37	64.89	32.95	30.25	1.100	0.510
Self-MM (AAAI'21)	70.87	69.30	68.74	66.75	33.65	30.82	1.065	0.522
TFR-Net (MM'21)	71.16	69.70	69.38	67.58	34.53	30.89	1.065	0.519
MMIM (EMNLP'21)	69.26	68.23	67.94	66.72	34.56	31.43	1.108	0.485
CENET (TMM'23)	70.43	68.19	66.79	64.01	33.55	30.54	1.069	0.511
TETFN (PR'23)	71.03	69.42	69.69	67.67	33.85	30.36	1.065	0.518
ALMT (EMNLP'23)	69.97	69.32	69.94	69.21	34.68	31.49	1.128	0.493
EMT-DLFR (TAC'24)	71.04	70.26	70.87	70.18	35.19	31.48	1.098	0.514
LNLN (NeurIPS'24)	70.09	69.50	70.01	69.49	34.28	31.20	1.124	0.501
P-RMF (ACL'25)	69.92	69.10	69.82	68.53	33.42	31.03	1.134	0.501
RECAP (AAAI'26)	70.18	69.48	70.08	69.29	33.82	30.75	1.100	0.505
SMCIR (AAAI'26)	71.51	70.89	71.44	70.73	36.59	32.60	1.072	0.528
MIDAS (TPAMI'26)	70.85	70.33	70.73	70.11	34.12	30.76	1.075	0.532
DREAM	**73.37**±1.04	**72.22**±0.94	**73.14**±1.04	**71.89**±1.08	**37.31**±1.87	**33.41**±1.53	**1.018**±0.028	**0.561**±0.020

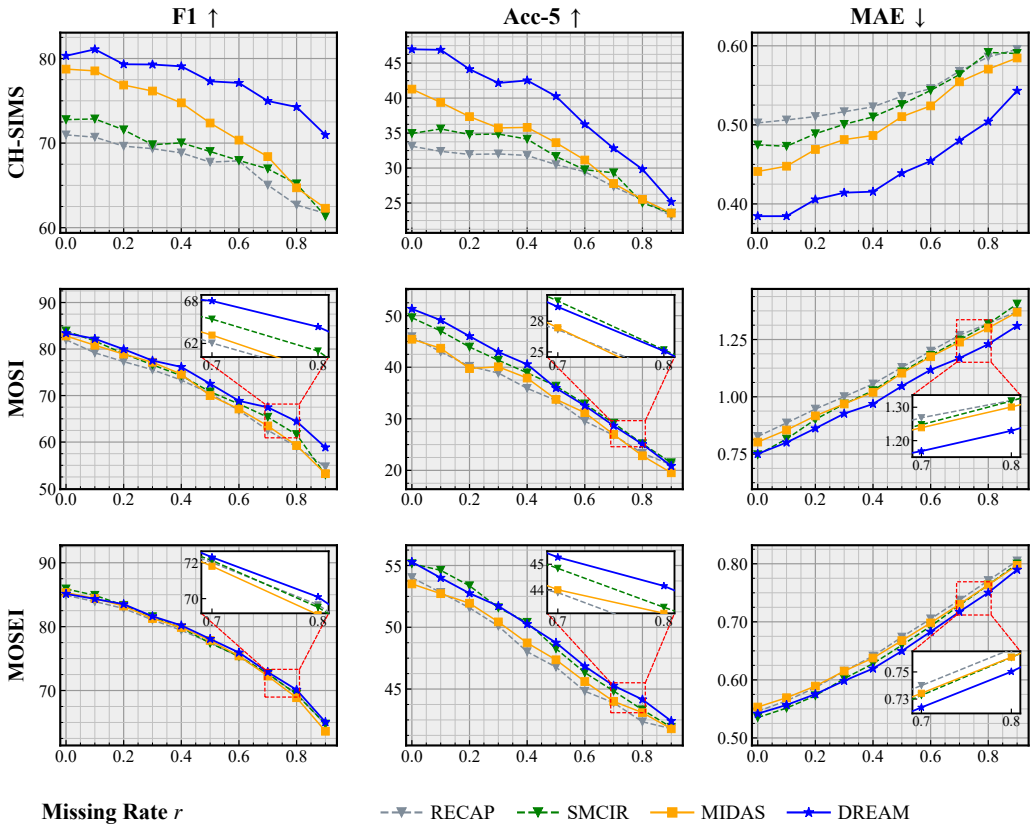

Figure 3: Performance across missing rates; English F1 excludes neutral targets. Insets enlarge $r = 0.7$–0.8; remaining metrics appear in Appendix A.1.

Appendix A extends this evaluation with Acc-2/Corr curves, complete per-rate scores, and sample-level polarity retention (Appendices A.1, A.2, and A.4).

Table 3: MOSEI results averaged over ten missing rates. DREAM: mean ± SD. Superscripts − and + exclude and include neutral targets.

Method	Acc-2^- ↑	Acc-2^+ ↑	$F1^-$ ↑	$F1^+$ ↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
MISA (MM'20)	72.73	76.44	66.82	69.95	38.70	37.90	0.791	0.497
Self-MM (AAAI'21)	73.66	76.60	69.73	73.53	46.46	45.78	0.701	0.496
TFR-Net (MM'21)	73.24	76.49	68.92	72.96	41.98	41.71	0.752	0.519
MMIM (EMNLP'21)	74.55	76.83	71.10	74.21	44.75	43.89	0.725	0.501
CENET (TMM'23)	73.85	76.94	68.84	72.93	45.70	44.88	0.719	0.531
TETFN (PR'23)	74.90	76.91	71.50	74.34	44.72	43.60	0.720	0.516
ALMT (EMNLP'23)	74.91	76.47	71.64	74.07	45.66	45.17	0.707	0.511
EMT-DLFR (TAC'24)	77.76	76.09	77.34	76.01	48.23	47.34	0.665	0.590
LNLN (NeurIPS'24)	76.78	74.90	76.04	74.88	46.61	45.84	0.678	0.581
P-RMF (ACL'25)	77.84	74.68	77.56	75.36	47.96	46.94	0.661	0.595
RECAP (AAAI'26)	77.57	75.44	77.17	75.79	47.62	46.79	0.664	0.589
SMCIR (AAAI'26)	77.91	76.05	77.51	76.36	48.98	47.96	0.653	0.586
MIDAS (TPAMI'26)	78.35	78.40	77.21	77.84	47.92	46.88	0.662	0.606
DREAM	**78.55**±0.16	**78.83**±0.29	**77.69**±0.39	**78.47**±0.01	**49.14**±0.56	**48.18**±0.61	**0.648**±0.007	**0.609**±0.009

Transferring to structured gaps. These additional tests erase contiguous segments or an entire modality. All models retain their random-missingness evaluation checkpoints, without reselection for structured conditions. DREAM achieves the lowest mean range-normalized MAE across the three datasets for every condition (Figure 4). Relative to the strongest comparator, it reduces this error by 8.0% and 4.7% under 50% and 90% span deletion, by 6.0% without text, and by 7.5% and 2.4% without audio and vision. These gains support transfer beyond random erasures. Appendix A.3 provides dataset-specific metrics and details of mask construction.

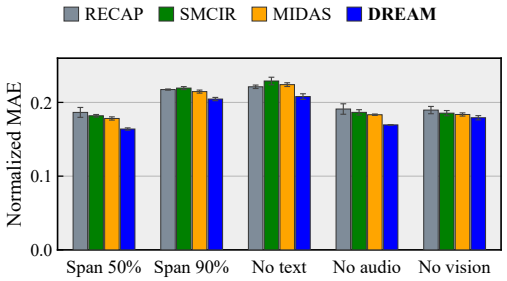

Figure 4: Structured-missingness MAE normalized by each dataset's sentiment range (lower is better), with equal dataset weights. Bars/error bars show mean/SD across three seeds.

4.3 Learning Task-Valued Local Evidence

Learning from unequal prediction damage. DREAM's regional supervision addresses a distinction that the missing rate alone misses: erasing equally sized regions can have different sentiment consequences. Matched CH-SIMS erasures produce opposite polarities in 9.3% of text pairs and 9.8% of visual pairs. Appendix B.2 quantifies these within-budget differences across modalities.

The learning controls in Table 4(a) link damage-guided learning and multiscale evidence to predictive gains. Removing ranking or damage-weighted assignment raises MOSI MAE from 1.018 to 1.048 or 1.042 and reduces every classification metric. Single-region supervision and removing learned lifting raise MAE to 1.057 and 1.037. Appendix B gives the exact interventions (B.1) and five-value analyses of both loss coefficients (B.4).

4.4 Reading Local Evidence beyond Summaries

Turning local evidence into predictive gains. The reading controls in Table 4(b) test how the learned evidence enters prediction. Summary-only exchange increases MAE to 1.045 and reduces nonneutral F1 by 1.18 points. Removing the task-value prior or receiver-local compatibility increases MAE to 1.042 or 1.035. DREAM leads seven of eight mean metrics; without compatibility, Acc-5 is slightly higher (37.36 versus 37.31).

Table 4: MOSI ablations: mean ± sample SD across three seeds, each averaged over ten missing rates. **Bold**: best mean; −/+ exclude/include neutral targets.

Variant	Acc-2$^-$ ↑	Acc-2$^+$ ↑	F1$^-$ ↑	F1$^+$ ↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
(a) Task-damage-guided representation learning								
w/o ranking	72.02±2.03	70.95±1.90	71.74±2.49	70.57±2.36	36.10±1.52	32.44±1.42	1.048±0.031	0.541±0.017
w/o assignment	72.71±0.32	71.76±0.40	72.65±0.43	71.60±0.49	35.66±1.19	31.86±1.15	1.042±0.005	0.550±0.005
Single-region	71.20±0.23	70.77±0.38	71.06±0.28	70.53±0.21	35.85±1.35	32.04±1.44	1.057±0.016	0.545±0.014
w/o lifting	72.05±0.88	71.28±0.80	71.95±0.97	71.08±0.90	36.35±1.04	32.59±1.08	1.037±0.019	0.544±0.013
(b) Summary-referenced local reading								
Summary-only	71.83±0.81	70.93±0.79	71.96±0.80	70.96±0.79	36.33±0.54	32.73±0.21	1.045±0.017	0.548±0.022
w/o value prior	72.84±0.97	71.89±0.69	72.91±0.94	71.87±0.67	35.61±0.32	32.01±0.33	1.042±0.011	0.547±0.014
w/o compatibility	72.40±0.74	71.55±0.30	72.28±0.69	71.33±0.16	**37.36**±0.76	33.22±0.90	1.035±0.019	0.553±0.004
DREAM	**73.37**±1.04	**72.22**±0.94	**73.14**±1.04	**71.89**±1.08	37.31±1.87	**33.41**±1.53	**1.018**±0.028	**0.561**±0.020

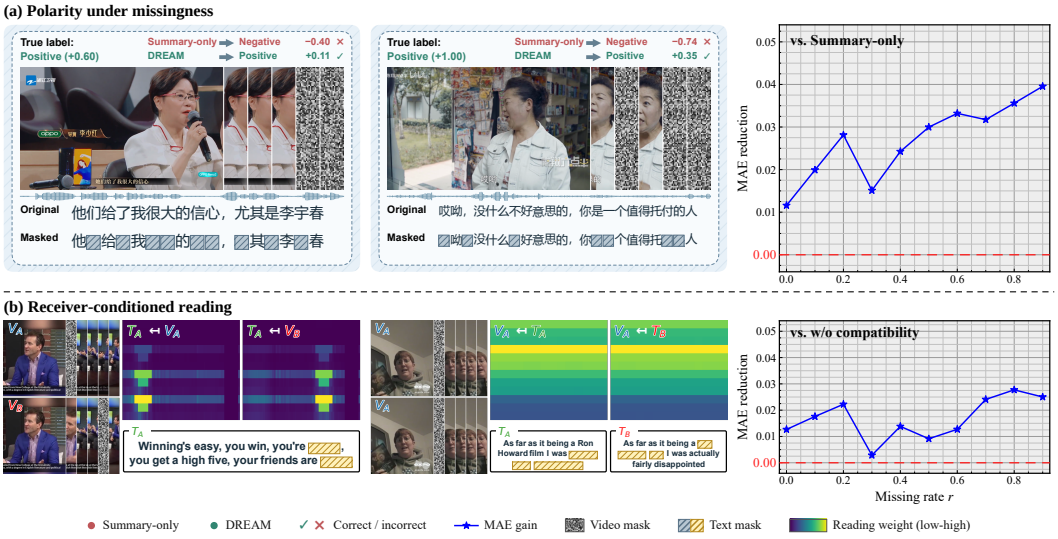

Figure 5: Prediction and local reading under missingness. (a) Selected CH-SIMS cases at $r = 0.5$ where DREAM corrects Summary-only polarity errors. (b) MOSEI weights with equal-budget source masks and a fixed receiver. T, V: text, vision; A, B: masks. Weights normalize within sources before outer gating; paired maps share a color scale. Right: control-minus-DREAM MAE on all 686 MOSI test clips, averaged across three seeds per rate; positive values favor DREAM.

Linking prediction with adaptive reading. Figure 5 pairs Chinese sentiment examples and local-reading maps with full-test MAE comparisons. DREAM reduces mean MAE at every tested rate versus Summary-only and the compatibility control, averaging 0.0269 and 0.0168, respectively. The maps illustrate reading as source availability changes. Appendix B.3 compares all seven controls across the full missing-rate range; Appendix C adds prediction examples and bidirectional reading cases from MOSI, MOSEI, and CH-SIMS.

5 Conclusion

DREAM learns task-valued multiscale evidence from the prediction damage of paired, equal-budget erasures without complete-view supervision, then reads local cues against global summaries according to each modality's context. MOSI, MOSEI, and CH-SIMS results show classification and regression gains under random and structured missingness. Future work will address natural sensor corruption, other multimodal tasks, and complete stream loss.

AI USE STATEMENT

AI was used to revise the wording of a very small number of sentences and to plan the layout of the appendix. All other data, figures, tables, and manuscript content were completed by the author personally.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix D.3 specify the model operators and training objectives. For reproducible evaluation, Appendix D.1 documents feature preprocessing, mask generation, metric conventions, checkpoint selection, and result aggregation. Architecture, optimization, software versions, hardware, and the timing procedure are provided in Appendices D.2 and D.5. Appendix A.2 reports all per-rate results with sample standard deviations, while Appendices A.3 and B.1 specify the structured-missingness and component-control experiments. We will publicly release the code to facilitate reproduction of our results.

REFERENCES

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208.

Jindi Bao, Jianjun Qian, Mengkai Yan, and Jian Yang. Progressive representation learning for multimodal sentiment analysis with incomplete modalities, 2026. arXiv:2603.09111.

Ying Cao, Penghui Zhao, Xinyu Qiao, Ningfan Zhan, and Xiaomei Zou. Data mining and uncertainty-aware with missing modalities for multimodal sentiment analysis. *Computers, Materials & Continua*, 88(3):23, 2026. doi: 10.32604/cmc.2026.084057.

Junsong Chen, Jiyuan Liu, Suyuan Liu, Wei Zhang, Ao Li, En Zhu, and Xinwang Liu. Sample-specific modality diagnosis and cross-modal enhancement for incomplete multimodal representations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):20154–20162, 2026. doi: 10.1609/aaai.v40i24.39102.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179.

Ian Covert, Tony Sun, James Y Zou, and Tatsunori Hashimoto. Locality alignment improves vision-language models. In *International Conference on Learning Representations*, pp. 83127–83165, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Inference-time dynamic modality selection for incomplete multimodal classification. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 116887–116913, 2026.

Osman Berke Guney, Ketan Suhaas Saichandran, Karim Elzokm, Ziming Zhang, and Vijaya B Kolachalama. Active feature acquisition via explainability-driven ranking. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 20748–20765, 2025.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, 2021. doi: 10.18653/v1/2021.emnlp-main.723.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 2020. doi: 10.1145/3394171.3413678.

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: ACL 2026*, pp. 5268–5288, 2026. doi: 10.18653/v1/2026.findings-acl.260.

Huiting Huang, Tieliang Gong, Kai He, Wen Wen, Weizhan Zhang, and Mengling Feng. Recovering coherent affective patterns: Addressing modality missing in multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(26):21957–21965, 2026. doi: 10.1609/aaai.v40i26.39349.

Atif Hussain, Yu Gu, Iftekharul Islam, Aamir Ali, He Zhang, and Chilato Chilato. Robust multimodal expert learning for sentiment analysis with missing modalities. *Scientific Reports*, 2026. doi: 10.1038/s41598-026-60943-7. Accepted manuscript, published online 31 July 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:18035–18043, 2025. doi: 10.1609/aaai.v39i17.33984.

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: EMNLP 2025*, pp. 11252–11267, 2025a. doi: 10.18653/v1/2025.findings-emnlp.602.

Yan Li, Yifei Xing, Xiangyuan Lan, Xin Li, Haifeng Chen, and Dongmei Jiang. AlignMamba: Enhancing multimodal Mamba with local and global cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24774–24784, June 2025b.

Han Lin, Xiu Tang, Huan Li, Wenxue Cao, Sai Wu, Chang Yao, Lidan Shou, and Gang Chen. T^2DR: A two-tier deficiency-resistant framework for incomplete multimodal learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8602–8616, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.452.

Ronghao Lin, Qiaolin He, Sijie Mai, Ying Zeng, Aolin Xiong, Li Huang, Yap-Peng Tan, and Haifeng Hu. CyIN: Cyclic informative latent space for bridging complete and incomplete multimodal learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 20032–20069, 2025b. doi: 10.52202/085713-0594.

Yanxi Luo, Shijin Wang, Zhongxing Xu, Yulong Li, Feilong Tang, and Jionglong Su. Confidence-aware self-distillation for multimodal sentiment analysis with incomplete modalities, 2025. arXiv:2506.01490.

Sijie Mai, Shiqin Han, and Haifeng Hu. Supervised attention mechanism for low-quality multimodal data. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21366–21386, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1084.

Phuong-Anh Nguyen, Tien Anh Pham, Duc-Trong Le, and Cam-Van Thi Nguyen. BALM: A model-agnostic framework for balanced multimodal learning under imbalanced missing rates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30246–30256, 2026.

Vittorio Pipoli, Alessia Saporita, Federico Bolelli, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. MissRAG: Addressing the missing modality challenge in multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3215–3224, 2025.

Licai Sun, Zheng Lian, Bin Liu, and Jianhua Tao. Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 15(1):309–325, 2024. doi: 10.1109/TAFFC.2023.3274829.

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. EmoMM: Benchmarking and steering MLLM for multimodal emotion recognition under conflict and missingness. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20351–20371, 2026. doi: 10.18653/v1/2026.findings-acl.1018.

Chuanqi Tao, Jiaming Li, Tianzi Zang, and Peng Gao. A multi-focus-driven multi-branch network for robust multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39:1547–1555, 2025. doi: 10.1609/aaai.v39i2.32146.

Di Wang, Xutong Guo, Yumin Tian, Jinhui Liu, LiHuo He, and Xuemei Luo. TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis. *Pattern Recognition*, 136: 109259, 2023a. doi: 10.1016/j.patcog.2022.109259.

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, 2023b. doi: 10.1109/TMM.2022.3183830.

Yuhua Wen, Yingying Zhou, Qifei Li, Yingming Gao, Zhengqi Wen, Jianhua Tao, and Ya Li. MIDAS: Mutual information disentanglement with uncertainty-aware fusion for incomplete multimodal sentiment analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026. doi: 10.1109/TPAMI.2026.3713694. Early access.

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

Kaixin Xu, NaiJin Liu, Yulin Kang, Tangyue Jin, Zixuan Yu, Wenxi Zhao, Yibei Liu, Qianle Zhang, Yangyang Wu, Mengying Zhu, and Meng Xi. Contrastive mixed prompt learning for incomplete multimodal sentiment analysis with unseen modality combination, 2026. arXiv:2608.20019.

Mingzheng Yang, Kai Zhang, Yuyang Ye, Yanghai Zhang, Runlong Yu, and Min Hou. Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 6803–6811, 2025. doi: 10.24963/ijcai.2025/757.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020. doi: 10.18653/v1/2020.acl-main.343.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797, 2021. doi: 10.1609/aaai.v35i12.17289.

Ziqi Yuan, Wei Li, Hua Xu, and Wenmeng Yu. Transformer-based feature reconstruction network for robust multimodal sentiment analysis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4400–4407, 2021. doi: 10.1145/3474085.3475585.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos, 2016. arXiv:1606.06259.

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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, 2023. doi: 10.18653/v1/2023.emnlp-main.49.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 55943–55974, 2024. doi: 10.52202/079017-1779.

Yujuan Zhang, Qing Li, Ziyu Li, Xiuxing Li, Zhuo Wang, Mengrui Xu, and Xia Wu. Active perceptual inference: A corticothalamic-inspired dynamic nested recurrent network for multimodal sentiment analysis with incomplete data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1787–1797, 2026.

Yuqing Zhang, Dongliang Xie, Dawei Luo, and Baosheng Sun. Emotional boundaries and intensity aware model for incomplete multimodal sentiment analysis. *Digital Signal Processing*, 160: 105023, 2025. doi: 10.1016/j.dsp.2025.105023.

Xinkui Zhao, Jinsong Shu, Yangyang Wu, Guanjie Cheng, Zihe Liu, Naibo Wang, Shuiguang Deng, Zhongle Xie, and Jianwei Yin. A robust incomplete multimodal low-rank adaptation approach for emotion recognition. *arXiv*, 2025. doi: 10.48550/arXiv.2507.11202.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22123–22138, 2025. doi: 10.18653/v1/2025.acl-long.1075.

Yitong Zhu, Yuxuan Jiang, Guanxuan Jiang, Bojing Hou, Peng Yuan Zhou, Ge Lin, and Yuyang Wang. QA-MoE: Towards a continuous reliability spectrum with quality-aware mixture of experts for robust multimodal sentiment analysis. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31689–31703, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1461.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Xiaoyue Zhang, Jiawen Deng, and Fuji Ren. CMAD: Correlation-aware and modalities-aware distillation for multimodal sentiment analysis with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4626–4636, 2025.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(3):2281–2289, 2026. doi: 10.1609/aaai.v40i3.37212.

Shihao Zou, Wei Wei, and Yongshuo Zhang. DEAR: Distributional error-aware reliability for robust multimodal sentiment analysis with missing modalities. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 30344–30361, 2026. doi: 10.18653/v1/2026.findings-acl.1517.

APPENDIX CONTENTS

A Extended Predictive Evaluation **14**
A.1 Missing-rate performance . . . 14
A.2 Complete prediction scores . . . 14
A.3 Transfer to structured missingness . . . 15
A.4 Polarity retention under severe missingness . . . 16
B Analysis of Evidence Learning and Reading **17**
B.1 Component Definitions . . . 17
B.2 Regional Prediction Damage . . . 18
B.3 Component Benefits across Missing Rates . . . 18
B.4 Sensitivity to the Loss Coefficients . . . 19
C Additional Prediction and Reading Examples **20**
C.1 Sentiment Prediction under Missing Inputs . . . 20
C.2 Bidirectional Reading on MOSEI . . . 21
C.3 Bidirectional Reading on CH-SIMS . . . 22
D Reproducibility **23**
D.1 Datasets and Evaluation Protocol . . . 23
D.2 Implementation Settings . . . 24
D.3 Operator and Objective Definitions . . . 24
D.4 Baseline Implementations . . . 26
D.5 Hardware and computational cost . . . 27

A EXTENDED PREDICTIVE EVALUATION

We extend Section 4.2 with additional metric curves, complete prediction scores, structured erasures, and sample-level polarity outcomes.

A.1 MISSING-RATE PERFORMANCE

Figure 6 extends the missing-rate comparison in Figure 3 to binary accuracy and correlation.

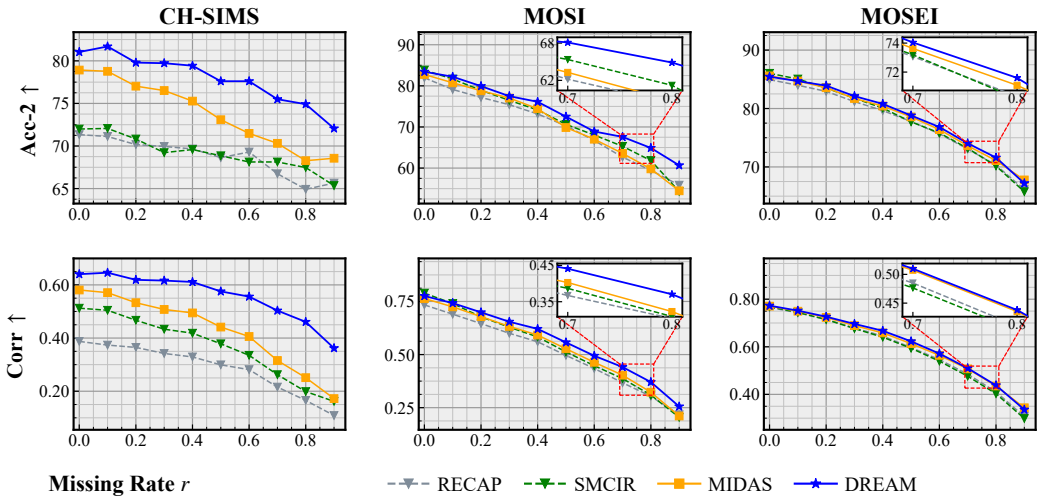

Figure 6: Acc-2 and Pearson correlation complement Figure 3. Curves show means across three seeds; English Acc-2 excludes neutral targets. White insets enlarge $r = 0.7$–0.8.

A.2 COMPLETE PREDICTION SCORES

Tables 5–7 report DREAM's full per-rate scores underlying the main benchmark averages. Metric definitions and aggregation follow Appendix D.1.

Table 5: DREAM on MOSI: mean ± SD across three seeds. Classification metrics are percentages. Superscripts −/+ exclude/include neutral targets.

Rate	Acc-2$^-$↑	Acc-2$^+$↑	F1$^-$↑	F1$^+$↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
0.0	83.44 ±0.92	81.97 ±0.93	83.41 ±0.87	81.89 ±0.92	51.31 ±1.16	44.70 ±1.08	0.751 ±0.021	0.776 ±0.002
0.1	82.21 ±0.84	80.86 ±0.55	82.20 ±0.75	80.79 ±0.46	49.13 ±2.20	42.91 ±1.40	0.800 ±0.026	0.742 ±0.012
0.2	79.93 ±1.04	78.57 ±1.19	79.96 ±0.99	78.53 ±1.15	46.02 ±3.51	40.09 ±2.53	0.862 ±0.040	0.699 ±0.018
0.3	77.49 ±1.23	76.24 ±1.39	77.52 ±1.22	76.20 ±1.39	42.95 ±3.06	37.51 ±2.26	0.926 ±0.026	0.655 ±0.013
0.4	76.12 ±1.10	74.88 ±1.49	76.15 ±1.15	74.83 ±1.56	40.57 ±3.72	36.10 ±2.93	0.969 ±0.026	0.621 ±0.012
0.5	72.51 ±1.53	71.48 ±1.75	72.53 ±1.59	71.41 ±1.84	35.96 ±2.56	32.36 ±2.29	1.046 ±0.033	0.557 ±0.017
0.6	68.90 ±1.19	67.93 ±1.58	68.86 ±1.27	67.78 ±1.77	32.56 ±2.48	29.84 ±2.33	1.118 ±0.024	0.494 ±0.021
0.7	67.58 ±1.77	66.47 ±1.58	67.47 ±1.40	66.25 ±1.24	28.67 ±0.66	26.43 ±0.60	1.169 ±0.036	0.441 ±0.040
0.8	64.89 ±2.59	63.80 ±1.98	64.44 ±1.85	63.21 ±1.20	25.07 ±1.68	23.81 ±1.60	1.230 ±0.025	0.370 ±0.032
0.9	60.67 ±3.54	59.96 ±2.49	58.86 ±3.43	58.00 ±3.01	20.85 ±0.96	20.36 ±0.81	1.309 ±0.045	0.256 ±0.042
Mean	73.37 ±1.04	72.22 ±0.94	73.14 ±1.04	71.89 ±1.08	37.31 ±1.87	33.41 ±1.53	1.018 ±0.028	0.561 ±0.020

Table 6 extends the rate-by-rate breakdown to MOSEI, retaining both neutral-target conventions.

Table 6: DREAM on MOSEI: mean ± SD across three seeds. Classification metrics are percentages. Superscripts −/+ exclude/include neutral targets.

Rate	Acc-2$^-$↑	Acc-2$^+$↑	F1$^-$↑	F1$^+$↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
0.0	85.45 ±0.44	84.67 ±0.45	85.13 ±0.54	84.60 ±0.38	55.29 ±1.15	53.44 ±1.19	0.542 ±0.017	0.773 ±0.004
0.1	84.69 ±0.20	84.20 ±0.42	84.33 ±0.17	84.10 ±0.32	54.00 ±1.29	52.30 ±1.34	0.557 ±0.015	0.752 ±0.007
0.2	83.91 ±0.35	83.48 ±0.23	83.51 ±0.39	83.36 ±0.18	52.75 ±1.05	51.31 ±1.23	0.575 ±0.010	0.727 ±0.008
0.3	82.12 ±0.32	82.03 ±0.09	81.63 ±0.38	81.86 ±0.10	51.72 ±1.32	50.50 ±1.44	0.598 ±0.014	0.696 ±0.008
0.4	80.79 ±0.30	80.89 ±0.40	80.19 ±0.23	80.65 ±0.33	50.25 ±0.84	49.16 ±0.98	0.620 ±0.010	0.667 ±0.013
0.5	78.81 ±0.67	79.15 ±0.65	78.08 ±0.69	78.85 ±0.60	48.75 ±0.72	47.86 ±0.85	0.650 ±0.002	0.623 ±0.011
0.6	76.83 ±0.56	77.28 ±0.39	75.94 ±0.72	76.92 ±0.41	46.84 ±0.35	46.18 ±0.40	0.683 ±0.008	0.571 ±0.017
0.7	74.04 ±0.14	75.08 ±0.59	72.89 ±0.42	74.56 ±0.39	45.27 ±0.38	44.82 ±0.39	0.717 ±0.002	0.510 ±0.008
0.8	71.60 ±0.46	72.77 ±1.04	70.12 ±0.82	72.09 ±0.32	44.15 ±0.34	43.89 ±0.38	0.750 ±0.004	0.438 ±0.013
0.9	67.21 ±0.87	68.74 ±3.30	65.07 ±1.78	67.71 ±1.27	42.41 ±0.42	42.35 ±0.40	0.790 ±0.003	0.336 ±0.017
Mean	78.55 ±0.16	78.83 ±0.29	77.69 ±0.39	78.47 ±0.01	49.14 ±0.56	48.18 ±0.61	0.648 ±0.007	0.609 ±0.009

Table 7 gives the CH-SIMS breakdown, including three- and five-class accuracy alongside binary polarity and regression.

Table 7: DREAM on CH-SIMS: mean ± SD across three seeds. Classification metrics are percentages.

Rate	Acc-2↑	F1↑	Acc-3↑	Acc-5↑	MAE↓	Corr↑
0.0	81.04 ±2.16	80.32 ±2.47	69.07 ±2.07	46.97 ±0.88	0.384 ±0.014	0.640 ±0.028
0.1	81.69 ±0.88	81.09 ±1.16	68.86 ±1.55	46.90 ±0.50	0.384 ±0.007	0.646 ±0.016
0.2	79.79 ±1.41	79.32 ±1.33	67.18 ±0.22	44.13 ±1.75	0.406 ±0.004	0.619 ±0.025
0.3	79.72 ±0.77	79.29 ±0.51	65.72 ±1.03	42.16 ±0.77	0.414 ±0.001	0.616 ±0.006
0.4	79.43 ±1.53	79.08 ±1.15	65.43 ±1.74	42.52 ±1.91	0.415 ±0.003	0.611 ±0.019
0.5	77.61 ±2.54	77.31 ±2.13	64.11 ±2.32	40.26 ±1.97	0.439 ±0.010	0.575 ±0.028
0.6	77.61 ±1.64	77.11 ±1.00	63.38 ±1.25	36.25 ±1.75	0.454 ±0.005	0.556 ±0.020
0.7	75.49 ±1.90	74.98 ±1.35	61.34 ±1.54	32.82 ±2.16	0.480 ±0.007	0.503 ±0.024
0.8	74.91 ±2.15	74.27 ±1.60	60.61 ±2.69	29.83 ±1.24	0.504 ±0.019	0.461 ±0.035
0.9	72.06 ±3.34	70.97 ±2.04	53.97 ±2.86	25.17 ±0.95	0.543 ±0.017	0.362 ±0.038
Mean	77.94 ±1.75	77.37 ±1.39	63.97 ±1.38	38.70 ±0.82	0.442 ±0.008	0.559 ±0.021

A.3 Transfer to structured missingness

DREAM has higher mean F1 and lower MAE than all three comparators in every evaluated MOSI condition (Table 8). On CH-SIMS, removing text gives DREAM an F1 of 74.55 and MAE of 0.515, compared with F1 of 56.82–59.42 and MAE of 0.580–0.594 for the comparators. Vision removal is the most demanding single-modality condition for CH-SIMS polarity and accounts for the small cross-dataset Acc-2/F1 gap; even there, DREAM attains the lowest MAE on every dataset.

These feature-level erasure tests use full test splits, keeping each checkpoint fixed across conditions without adaptation to structured missingness. Span masks relocate each clip's random deletions into a contiguous span per modality, preserving deletion counts. Single-modality removal masks all content in one modality while the others have 50% random missingness. Special tokens and padding stay fixed. All comparators share DREAM's actual masks, sample IDs, labels, and order; the 50% and 90% random conditions also verify agreement with native checkpoint inference. Metrics are recomputed from saved predictions and summarized by the mean and sample SD across three seeds. Baseline implementation and training settings are detailed in Appendix D.4.

Table 8: Structured missingness: mean ± sample SD across three seeds. F1 is in percent; English F1 excludes neutral targets. Bold marks the best displayed mean (ties at displayed precision). SMCIR on CH-SIMS is a project adaptation (Appendix D.4).

	RECAP		SMCIR		MIDAS		DREAM	
Condition	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓	F1↑	MAE↓
MOSI								
Span, 50%	69.57 ±1.08	1.116 ±0.020	70.23 ±1.27	1.095 ±0.026	69.54 ±1.55	1.102 ±0.028	**72.37** ±1.92	**1.054** ±0.019
Span, 90%	53.73 ±3.39	1.362 ±0.029	51.35 ±3.57	1.403 ±0.050	54.05 ±4.01	1.359 ±0.021	**58.75** ±2.68	**1.303** ±0.023
No text	49.52 ±3.46	1.406 ±0.030	34.82 ±16.87	1.512 ±0.085	44.78 ±6.09	1.431 ±0.043	**57.10** ±4.99	**1.360** ±0.060
No audio	70.44 ±1.75	1.129 ±0.010	70.76 ±1.06	1.108 ±0.026	70.11 ±0.36	1.103 ±0.020	**72.04** ±1.57	**1.050** ±0.034
No vision	70.44 ±1.75	1.128 ±0.011	70.76 ±1.09	1.108 ±0.026	70.39 ±0.08	1.091 ±0.031	**70.95** ±1.32	**1.055** ±0.035
MOSEI								
Span, 50%	77.83 ±0.50	0.653 ±0.007	**78.49** ±0.81	0.640 ±0.001	78.22 ±0.70	0.652 ±0.004	78.39 ±0.09	**0.636** ±0.007
Span, 90%	65.78 ±0.17	0.798 ±0.005	**66.12** ±0.28	0.788 ±0.005	63.98 ±1.34	0.792 ±0.014	65.61 ±1.36	**0.782** ±0.006
No text	55.28 ±2.46	0.835 ±0.007	56.74 ±2.16	0.830 ±0.003	52.22 ±3.67	**0.819** ±0.008	**58.90** ±4.00	0.837 ±0.024
No audio	77.49 ±0.77	0.674 ±0.010	77.35 ±0.92	0.661 ±0.002	78.02 ±1.33	0.666 ±0.008	**78.15** ±0.69	**0.650** ±0.002
No vision	77.23 ±0.53	0.676 ±0.009	**77.52** ±1.10	0.662 ±0.002	76.93 ±1.53	0.684 ±0.014	77.14 ±0.94	**0.661** ±0.001
CH-SIMS								
Span, 50%	67.76 ±5.18	0.529 ±0.039	70.38 ±0.88	0.513 ±0.013	73.50 ±1.21	0.485 ±0.005	**79.14** ±0.73	**0.420** ±0.011
Span, 90%	63.60 ±2.13	0.585 ±0.012	62.49 ±2.11	0.586 ±0.008	64.56 ±0.71	0.571 ±0.004	**70.45** ±1.11	**0.532** ±0.007
No text	58.61 ±3.10	0.580 ±0.009	59.42 ±0.51	0.593 ±0.002	56.82 ±0.00	0.594 ±0.007	**74.55** ±0.97	**0.515** ±0.004
No audio	66.32 ±8.24	0.545 ±0.043	67.96 ±0.88	0.529 ±0.022	72.25 ±1.74	0.510 ±0.009	**76.88** ±2.65	**0.451** ±0.009
No vision	67.75 ±4.93	0.536 ±0.031	68.88 ±2.25	0.522 ±0.017	**72.34** ±1.59	0.510 ±0.009	70.47 ±0.86	**0.503** ±0.008

A.4 Polarity retention under severe missingness

Overall, correct-polarity gains outnumber losses at both 50% and 90% missingness, averaged across three seeds (Figure 7). At 90%, DREAM's net accuracy gains over RECAP are 4.83, 1.10, and 6.42 percentage points on MOSI, MOSEI, and CH-SIMS. The overlap analysis includes 656 nonneutral MOSI clips, 3,634 nonneutral MOSEI clips, and all 457 CH-SIMS clips, without correctness filtering, and pairs models trained with the same seed.

Among the 530 MOSI clips both DREAM and Summary-only classify correctly before masking, DREAM retains correct polarity in 70 cases lost by Summary-only at $r = 0.9$, versus 44 in the reverse direction. Appendix C.1 illustrates selected successes at $r = 0.5$.

Table 9: F1 gains ↑ at 90% missing (pp).

Dataset	1111	1112	1113	Mean
MOSI	+6.03	+5.86	+5.06	+5.65
MOSEI	+1.47	+1.04	+1.81	+1.44
CH-SIMS	+7.47	+8.68	+9.86	+8.67

DREAM also improves F1 over MIDAS for each of the three seeds on every dataset at $r = 0.9$ (Table 9). The gains span 5.06–6.03 percentage points on MOSI, 1.04–1.81 on MOSEI, and 7.47–9.86 on CH-SIMS.

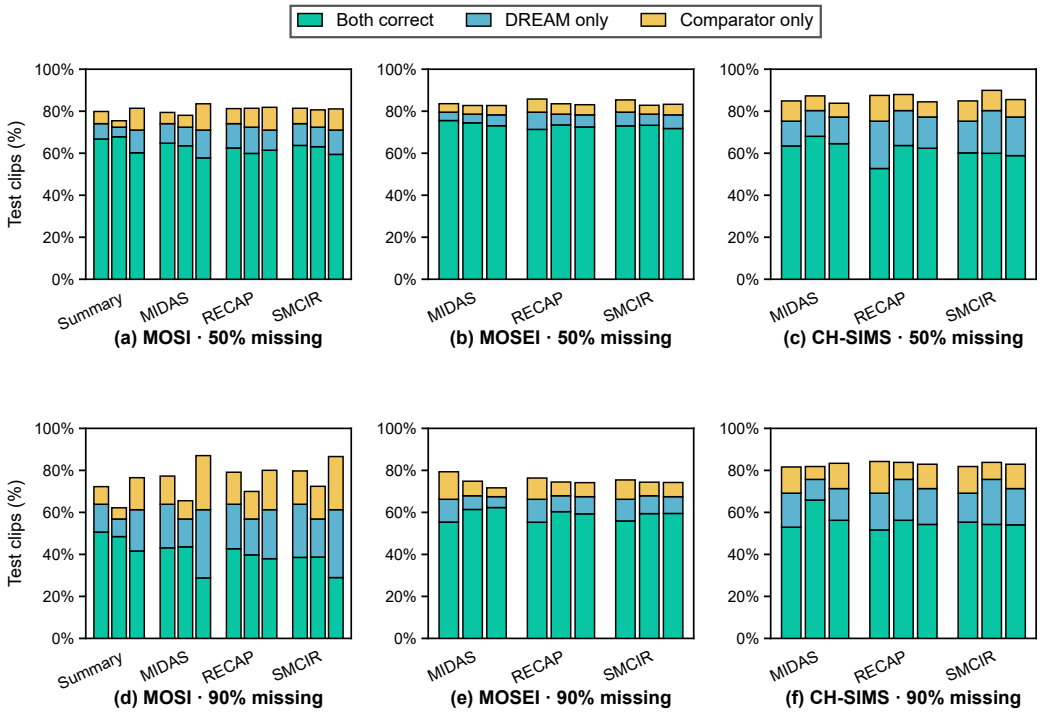

Figure 7: Sample-level polarity overlap across datasets and missing rates. Comparators are MIDAS, RECAP, and SMCIR on all three datasets, with Summary-only additionally shown on MOSI. Each group shows three seeds in the order given in Section 4.1. Green: both correct; blue: DREAM alone correct; yellow: comparator alone correct. Unfilled space to 100% denotes joint errors: total bar height is the union of correct predictions, not DREAM accuracy. English sets exclude zero targets; CH-SIMS uses its original binary convention.

B Analysis of Evidence Learning and Reading

The analyses below extend Sections 4.3 and 4.4: component definitions specify the interventions, matched erasures establish unequal regional damage, rate-resolved ablations test both modules, and coefficient sweeps characterize their training settings.

B.1 Component Definitions

This section specifies the interventions behind Table 4, complementing the learning and reading analyses in Sections 4.3 and 4.4. All controls use the MOSI training, selection, and aggregation protocol in Appendix D.1, with matching test examples and masks.

Evidence-learning controls. Let 0 denote the original incomplete input and A and B its additionally erased views. The ranking control disables $(A, 0)$, $(B, 0)$, and (A, B) comparisons while retaining damage-weighted assignment; the assignment control retains the rankings. Single-region supervision keeps $(A, 0)$ and assignment for A, changing both the comparison set and assignment views. The lifting control replaces the learned low-frequency correction with its smoothed average.

Evidence-reading controls. Summary-only disables local reading during both training and evaluation. The value-prior control uses a uniform prior over valid scale–position pairs; learned task values continue to guide decomposition and receive regional supervision. The compatibility control zeros the local cosine term while retaining the task-value prior, summary competition, and receiver-conditioned outer source selection.

Per-rate prediction comparison. The right panels of Figure 5 use all 686 MOSI test clips, including neutral targets, and unclipped continuous predictions. For each matched run and rate we subtract DREAM's MAE from the control's MAE, then average across three seeds. This rate-resolved view complements the complete classification and regression comparison in Table 4.

B.2 Regional Prediction Damage

The paired-region motivation in Section 4.3 can be examined directly by holding the deletion budget fixed. From the first 128 CH-SIMS test clips at $r = 0.5$, matched erasures across three seeds and three scales yield 1,583 active pairs: 538 text, 841 audio, and 204 visual. Each pair starts from the same incomplete input and preserves valid width and erased-position count.

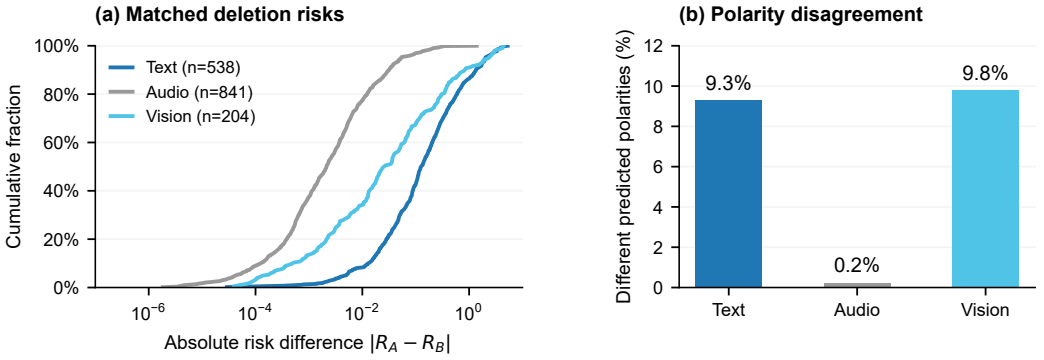

Figure 8: Prediction damage under matched regional erasures on CH-SIMS. (a) Cumulative distributions of $|R_A - R_B|$ on a log axis, with $R_U = \mathcal{R}(\tilde{x}^{(-U)}, y)$. (b) Opposite-polarity frequency. Active pairs include repeated clips across runs and scales; modality counts reflect different sequence lengths and candidate coverage.

Damage varies within a fixed budget. Median absolute risk gaps for text/audio/vision are 0.1272/0.0021/0.0263; their 90th percentiles are 1.3335/0.0316/0.8253. The distributions reveal variation in the consequence of removing different content, providing the training signal used by the regional objectives.

These within-budget differences supply a task-specific learning signal: the regional objectives distinguish which erasures matter for the current prediction, while matching the amount of removed content. The resulting contribution to prediction is tested by the ranking, assignment, single-region, and lifting controls in Table 4; Appendix B.3 resolves their effects across missing rates.

B.3 Component Benefits across Missing Rates

The learning and reading gains extend across the missing-rate range. Figure 9 resolves all seven controls from Table 4 into their ten evaluation rates. At every rate, the complete model achieves lower mean MAE than each control, yielding 70 positive comparisons across three seeds.

Evidence learning under increasing information loss. Single-region supervision has the largest average MAE gap among the learning controls. The contribution of learned lifting grows from 0.0072 at complete input to 0.0488 at 90% missingness; damage-weighted assignment similarly grows from 0.0160 to 0.0430. The rate-resolved results connect multiscale representation and regional supervision to retaining predictive information as observations disappear.

Complementary ingredients of local reading. Summary-only, a uniform task-value prior, and removing local compatibility each increase mean MAE throughout the range. The prior comparison tests the use of learned task value in reading, while the compatibility comparison tests content matching with the receiver. Together with the precise controls in Appendix B.1, these results support combining task value, local compatibility, and summary-referenced exchange.

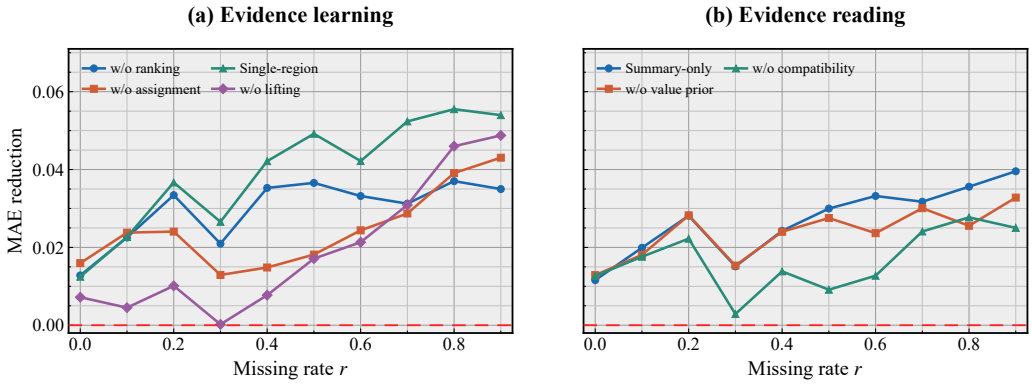

Figure 9: Component contributions across missing rates on the complete MOSI test set. Curves show control-minus-DREAM MAE averaged across three seeds: (a) four evidence-learning controls; (b) three evidence-reading controls. Each control uses matching test clips and masks, with one validation-selected checkpoint per run for all rates. Positive values indicate a benefit from the corresponding DREAM design.

B.4 Sensitivity to the Loss Coefficients

The two coefficients in Eq. 7 balance the regional auxiliary objectives and sentiment intensity. We evaluate each at $\{1/4, 1/2, 1, 2, 4\}$ times its fixed default, holding the other constant. The tested auxiliary settings are $\lambda_{\text{aux}} \in \{0.05, 0.10, 0.20, 0.40, 0.80\}/7$, and the intensity settings are $\lambda_I \in \{0.125, 0.25, 0.50, 1.00, 2.00\}$. The defaults are $0.2/7$ and 0.5.

The scans characterize the fixed defaults on the MOSI benchmark used during model development, following Appendix D.1. Both sweeps share the default models. The fixed defaults yield the highest mean F1 (73.14%) and lowest MAE (1.018) in both scans. Across the tested range, F1 spans 70.42–73.14% for the auxiliary coefficient and 70.92–73.14% for the intensity coefficient. Figure 10 displays the response around the defaults; Table 10 adds all eight metrics.

Table 10: MOSI five-point sensitivity, reported as mean ± sample SD across three seeds, each averaged over ten missing rates. Superscripts −/+ exclude/include neutral targets.

Multiple	Acc-2$^-$ ↑	Acc-2$^+$ ↑	F1$^-$ ↑	F1$^+$ ↑	Acc-5↑	Acc-7↑	MAE↓	Corr↑
Auxiliary coefficient: fixed default $\lambda_{\text{aux}} = 0.2/7$								
1/4	71.32±1.34	70.68±1.20	70.91±1.85	70.16±1.70	35.58±1.45	31.83±1.12	1.053±0.009	0.548±0.014
1/2	70.73±1.11	70.32±0.87	70.42±1.41	69.91±1.19	34.81±1.04	31.05±0.84	1.064±0.015	0.533±0.008
1 (default)	73.37±1.04	72.22±0.94	73.14±1.04	71.89±1.08	37.31±1.87	33.41±1.53	1.018±0.028	0.561±0.020
2	72.44±1.22	71.65±0.91	72.41±1.23	71.52±0.93	36.04±1.00	32.10±0.94	1.043±0.028	0.550±0.011
4	72.11±0.83	71.32±0.66	72.09±0.92	71.20±0.76	36.24±0.70	32.24±0.87	1.043±0.009	0.551±0.012
Intensity coefficient: fixed default $\lambda_I = 0.5$								
1/4	72.73±1.32	71.54±0.71	72.52±1.51	71.23±0.90	36.77±1.40	32.95±1.08	1.028±0.023	0.561±0.006
1/2	71.06±1.21	70.53±1.03	70.92±1.57	70.30±1.39	35.39±1.76	31.65±1.51	1.057±0.010	0.542±0.015
1 (default)	73.37±1.04	72.22±0.94	73.14±1.04	71.89±1.08	37.31±1.87	33.41±1.53	1.018±0.028	0.561±0.020
2	71.62±0.89	70.84±1.07	71.63±0.90	70.75±1.09	35.51±1.24	31.95±0.71	1.060±0.006	0.532±0.013
4	71.88±1.20	70.81±0.67	71.85±1.29	70.68±0.72	35.86±1.57	32.09±1.61	1.049±0.019	0.541±0.012

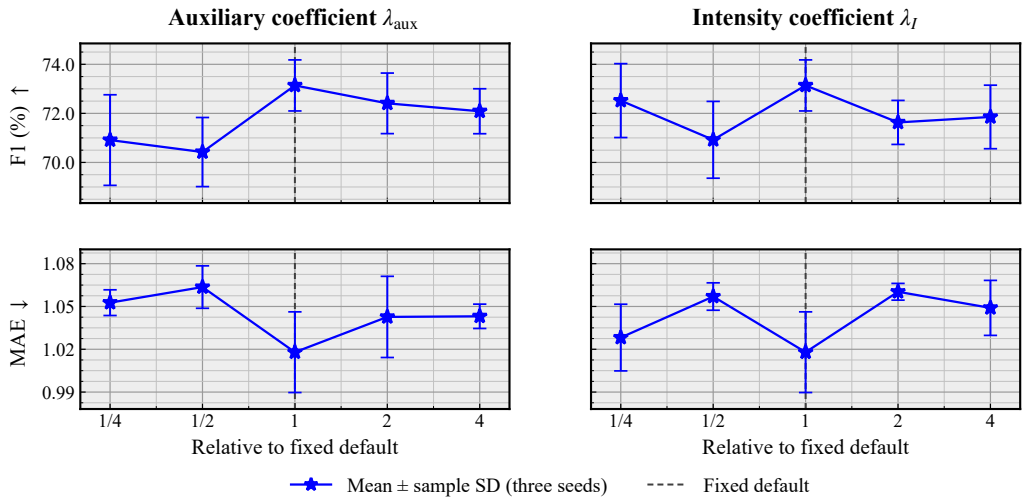

Figure 10: Sensitivity to auxiliary and intensity supervision on MOSI. Horizontal axes give multiples of the fixed defaults, marked by dashed lines. Points and bars show means and sample SD across three seeds, each averaged over missing rates; F1 excludes neutral targets.

C Additional Prediction and Reading Examples

C.1 Sentiment Prediction under Missing Inputs

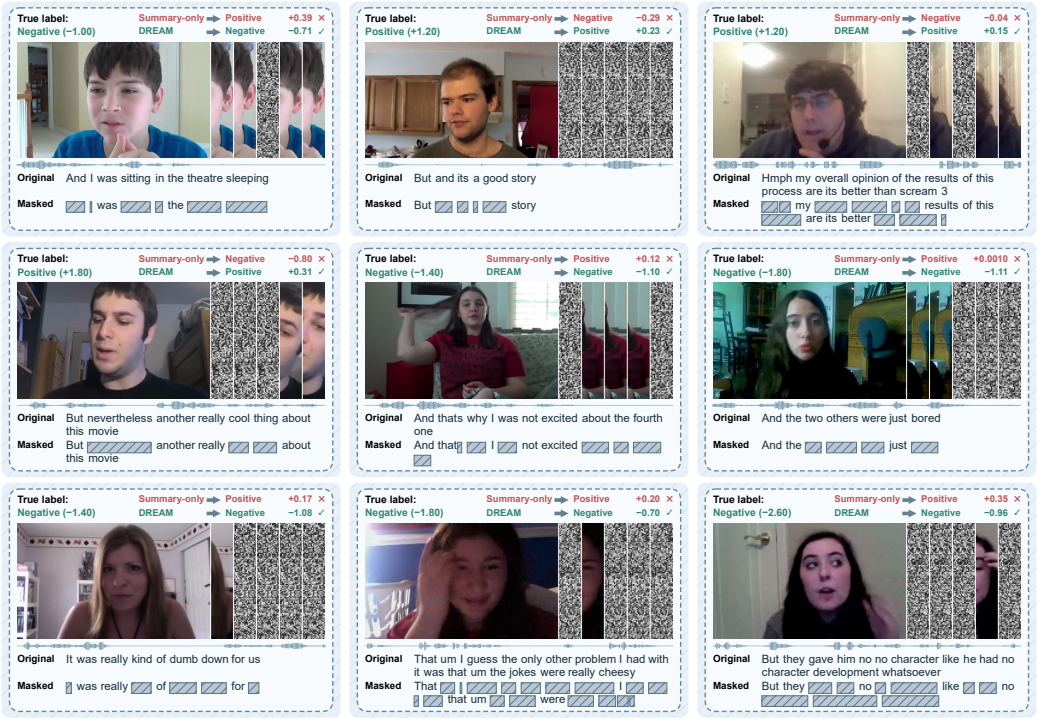

Figure 11: Eighteen MOSI polarity-preservation cases extending Figure 5 ($r = 0.5$). Both models are correct before masking; only DREAM is correct afterwards. Headers show targets and predictions (green: correct/reference; red: incorrect). Video strips mark feature erasures, waveforms show original audio, and hatching marks masked WordPieces.

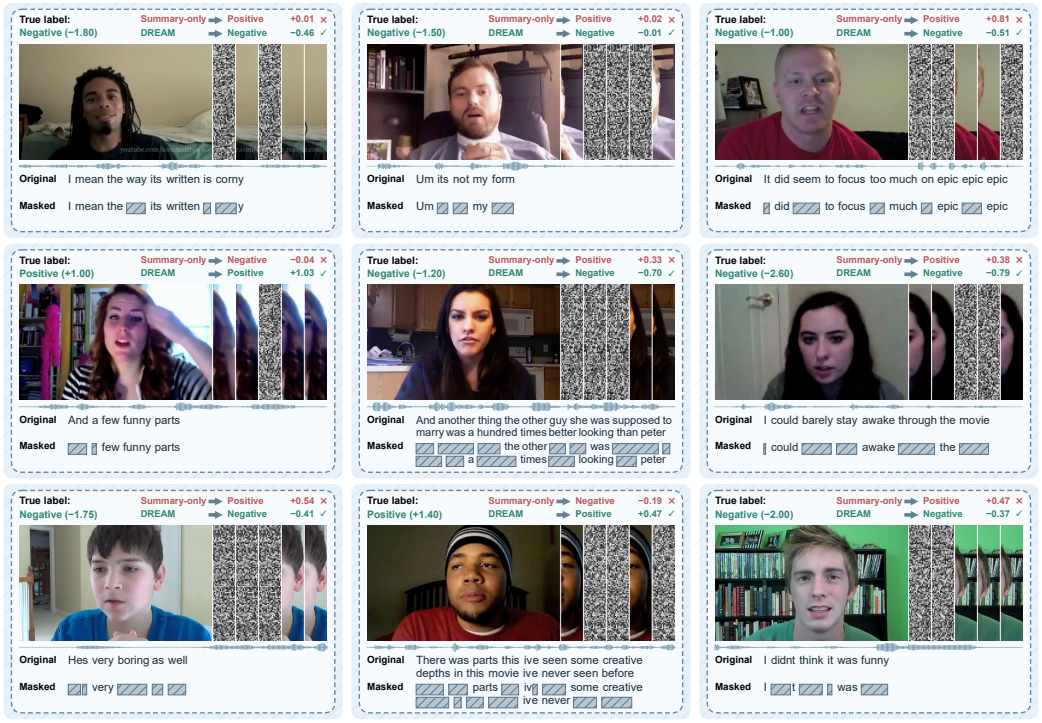

Figure 11 (continued).

C.2 Bidirectional Reading on MOSEI

In the following grids, T, V denote text and vision and A, B identify masks. Left columns fix the text receiver while varying visual-source masks; right columns reverse the roles. Each pair preserves source deletion counts, receiver representations, and audio. Weights are normalized within each source before outer source/null gating. The clips and masks were selected for varied, readable content before weight export. Masked positions can retain contextual states and receive reading weight.

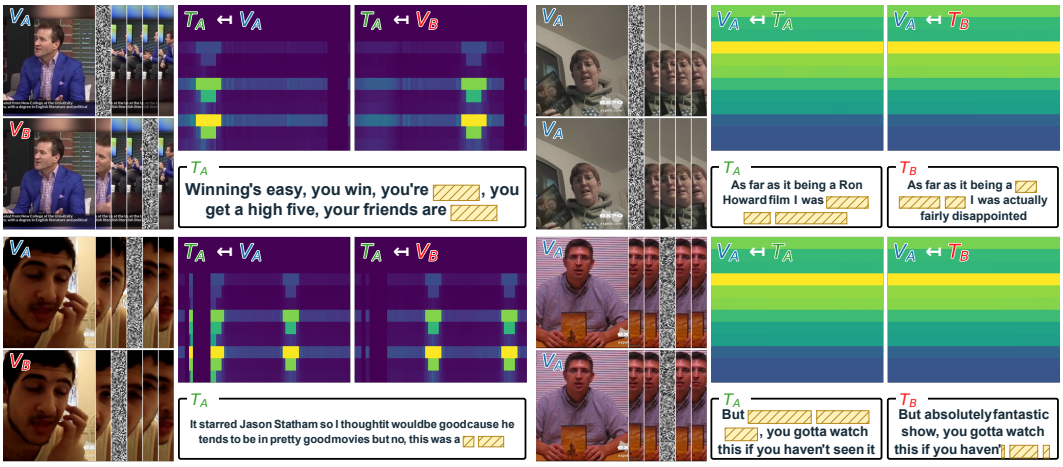

Figure 12: Bidirectional reading in ten distinct MOSEI videos. Arrows denote source-to-receiver reading. Native maps stack four heads and three scales, with source positions as columns and a shared linear color scale within each pair. Video strips show five equally spaced feature-mask samples; static texture marks erased features and gold hatching marks masked text.

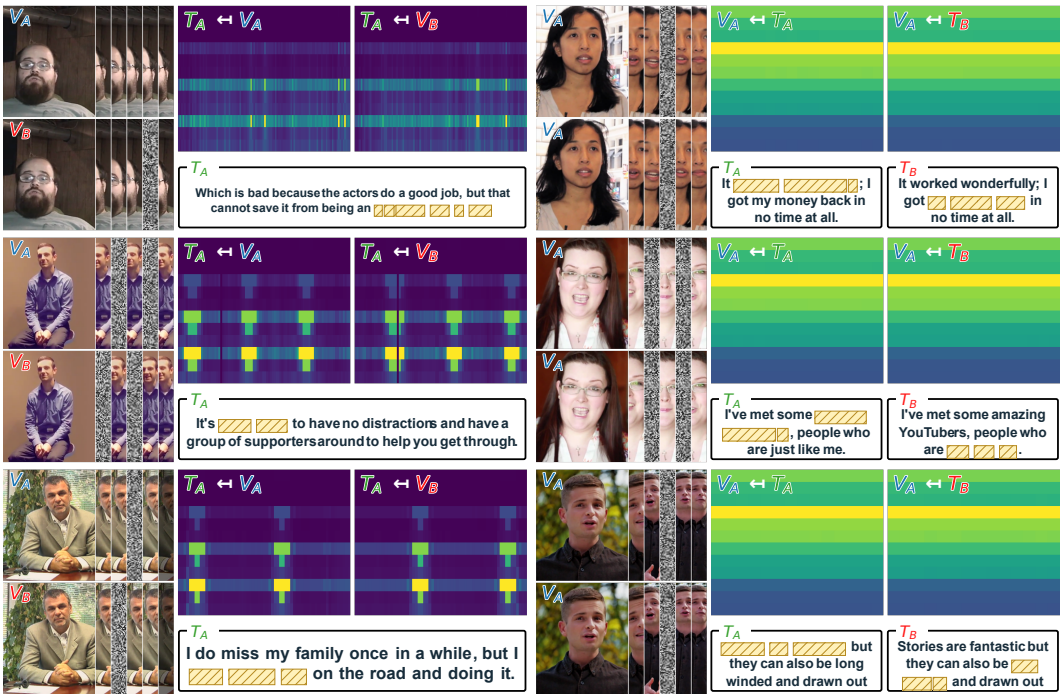

Figure 12 (continued).

C.3 Bidirectional Reading on CH-SIMS



Figure 13: Bidirectional reading in ten distinct CH-SIMS videos, with the conventions of Figure 12. Left: $T_A \leftarrow V_A/V_B$; right: $V_A \leftarrow T_A/T_B$. These additional Chinese examples extend the same source-availability intervention beyond the MOSEI cases in Figure 5.



Figure 13 (continued).

D REPRODUCIBILITY

This section supplies the evaluation and implementation details supporting Section 4.1, followed by the expanded operators of Section 3, baseline configurations, and inference measurements.

D.1 DATASETS AND EVALUATION PROTOCOL

Data and features. We use the unaligned feature sets and splits adopted by MIDAS (Wen et al., 2026) for MOSI (Zadeh et al., 2016), MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020). Table 11 lists padded input lengths and feature widths; text includes special tokens. Text uses fine-tuned BERT (Devlin et al., 2019). CH-SIMS visual channels are divided by training-set RMS values from 30,339 nonzero frames in 1,368 clips. No mean subtraction is applied, preserving zeros. DREAM uses multimodal sentiment labels, without CH-SIMS unimodal supervision.

Table 11: Dataset splits and input dimensions.

	MOSI	MOSEI	CH-SIMS
Train	1,284	16,326	1,368
Valid	229	1,871	456
Test	686	4,659	457
Label range	$[-3, 3]$	$[-3, 3]$	$[-1, 1]$
Input length			
Text	50	50	39
Audio	375	500	400
Vision	500	500	55
Feature width			
Text	768	768	768
Audio	5	74	33
Vision	20	35	709

Missing observations. A valid position is retained when a uniform draw exceeds the erasure probability r. Erased tokens become `[UNK]` (ID 100); erased audio/visual vectors become zero. Special tokens `[CLS]`/`[SEP]` and padding stay fixed. Training draws a rate per clip and modality, then randomly sets half the clips' rates to zero in each modality. Initial rates follow the uniform distribution $\mathcal{U}(0, 0.9)$; later draws use $\mathcal{U}(0, 1)$ on MOSI and $\mathcal{U}(0, 0.9)$ on MOSEI/CH-SIMS. The loader refreshes rates and masks whenever index zero is visited in shuffled order, using mask seed 9001. Evaluation uses one nominal rate for all modalities, regenerating each modality's uniform array from the run seed. Thresholding the same draws nests masks across rates. Rates are erasure probabilities, not exact per-clip counts.

Metrics and neutral labels. For English Acc-2 and F1, superscripts $-$ and $+$ exclude and include zero-valued targets. The former removes targets with $y = 0$ and uses > 0 for positive predictions and labels; the latter retains all targets and uses ≥ 0. F1 is support-weighted. English Acc-5/Acc-7 round predictions and labels after clipping to $[-2, 2]$/$[-3, 3]$; MAE and Pearson correlation use the un-clipped scores. CH-SIMS clips both scores to $[-1, 1]$ before all metrics. Its binary split is $[-1, 0]$ versus $(0, 1]$; Acc-3 uses cut points $-0.1, 0.1$, and Acc-5 uses $-0.7, -0.1, 0.1, 0.7$, assigning boundary values to the lower interval. Classification scores are percentages; MAE retains the dataset's sentiment units.

Selection and aggregation. Each run selects the earliest minimum validation loss after epoch one at $r = 0.5$: MSE-based loss on MOSI/MOSEI, MAE on CH-SIMS. English loss retains $BK^{-1}\sum_{k=1}^{K}\mathrm{MSE}_k$, where MSE_k is the mean squared error of batch k, K is the batch count, and nominal batch size $B = 32$ also applies to a short final batch. One checkpoint evaluates all metrics at $r \in \{0.0, 0.1, \ldots, 0.9\}$. Our ordinary evaluations, all seven ablations, and structured tests use the three seeds specified in Section 4.1. Structured tests keep each checkpoint fixed across conditions. For ten-rate results, we average the ten scores within each seed, then report the mean and sample SD across three seeds. Per-rate and structured-condition SDs use the three seed-level scores for that condition. Appendix D.4 gives baseline-specific training and selection settings.

D.2 Implementation Settings

DREAM uses bidirectional GRUs (Cho et al., 2014) for audio and vision, with fine-tuned `bert-base-uncased` on MOSI/MOSEI and `bert-base-chinese` on CH-SIMS. Table 12 gives model and training settings.

Table 12: Implementation settings for DREAM. Shared settings apply to all three datasets.

Shared model settings		**Shared optimization settings**	
Hidden size	128	Optimizer	Adam
Number of scales	3	Batch size	32
Attention heads	4	BERT learning rate	10^{-5}
GRU layers	2	Learning-rate schedule	Cosine decay
GRU units per direction	64	Linear warmup	10% of epochs
GRU dropout	0	Precision	FP32

Dataset-specific settings

Setting	MOSI	MOSEI	CH-SIMS
BERT backbone	Uncased English		Chinese
Non-BERT learning rate	2×10^{-3}	1.5×10^{-3}	2×10^{-3}
Training epochs	100	20	100
Neutral polarity weight	1.0	0.6	1.0

Optimization. Adam (Kingma & Ba, 2015) uses $(\beta_1, \beta_2) = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. Weight decay is 10^{-4} for BERT weights other than biases and LayerNorm parameters, zero for those excluded parameters, and 10^{-3} elsewhere. Learning rates are updated once per epoch, with linear warmup followed by cosine decay. The effective auxiliary and intensity coefficients in Eq. 7 are $0.2/7$ and 0.5, respectively; Appendix D.3 specifies the fixed auxiliary scaling used by the implementation.

Initialization. Lifting output layers start at zero; summary value and output projections start at identity, while local value projections are independently initialized. Appendix D.5 specifies software and hardware.

D.3 Operator and Objective Definitions

This section expands the operators in Section 3 and specifies the normalization and gradient paths of its auxiliary objectives.

Input representations and multiscale split. Text representations project the final BERT states to width d. For audio and vision, a linear–GELU–LayerNorm map φ_m gives $H_m = \varphi_m(\widetilde{x}_m) - \varphi_m(0)$, with padding positions zeroed. The split precedes the bidirectional GRU. Original padding validity $v_{m,\tau} \in \{0, 1\}$ equals one at nonpadding positions and is retained under feature erasure. It does not identify erased content. CH-SIMS visual features use the training-set channel RMS normalization specified in Appendix D.1.

For scale $s \in \{1, \ldots, S\}$, let $d_s = 2^{s-1}$ be the dilation. With tap offset $r \in \{-1, 0, 1\}$, a learned symmetric kernel has positive taps $k_{m,s,-1} = k_{m,s,1}$ and $\sum_{r=-1}^{1} k_{m,s,r} = 1$. The padding-normalized

average is

$$\overline{a}_{m,s,\tau} = \frac{\sum_{r=-1}^{1} k_{m,s,r} v_{m,\tau+rd_s} a_{m,s-1,\tau+rd_s}}{\sum_{r=-1}^{1} k_{m,s,r} v_{m,\tau+rd_s}}. \tag{8}$$

Indices outside the sequence are invalid, and zero-support positions return zero. As illustrated in Figure 14, learned lifting adds $L_{m,s}([a_{m,s-1}; \overline{a}_{m,s}; a_{m,s-1} - \overline{a}_{m,s}])$ to this average, yielding $a_{m,s}$. The residual is $\delta_{m,s} = a_{m,s-1} - a_{m,s}$, and its task-value score is $q_{m,s} = \sigma(\phi_{m,s}([H_m; a_{m,s}; \delta_{m,s}]))$. The maps $L_{m,s}$ and $\phi_{m,s}$ use LayerNorm followed by $3d \to d \to d$ and $3d \to d \to 1$ MLPs, respectively, with GELU between linear layers. Padding is zeroed after each stage.

The text summarizer $\mathcal{A}_t$ selects the classification token. Audio/visual summarizers apply the same two-layer BiGRU and LayerNorm to H_m, G_m^{seq}, and a zero sequence, subtract the corresponding zero-input response, and pool valid positions with learned content weights. The weights are computed separately for each input using shared parameters. Consequently $E_m = \mathcal{A}_m(H_m) - \mathcal{A}_m(G_m^{\text{seq}})$ is a difference of summaries, rather than an average of local detail vectors. Local memory excludes padding and text classification/separator tokens. Its validity includes currently masked positions: the reader has no oracle mask distinguishing observed from erased content.

Reader queries and keys use LayerNorm followed by a $2d \to d$ projection, partitioned into four heads $h \in \{1, 2, 3, 4\}$ with $d_h = d/4$. The reference/local maps $W_{\text{ref},j}^{(h)}$ and $W_{\text{loc},j}^{(h)}$ are the $d \to d_h$ head blocks of bias-free $d \to d$ projections. Invalid local entries have logit $-\infty$; empty support returns summary-reference probability one. The factors $\sqrt{d_h}$, S, and SN_j are fixed by head width and memory dimensions.

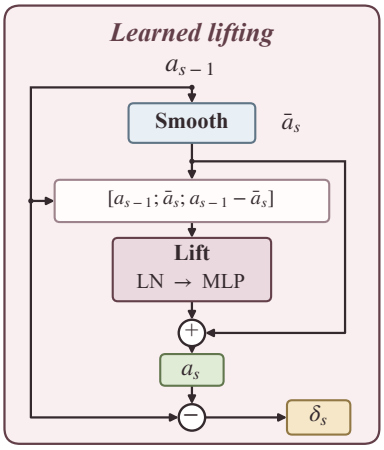

Figure 14: One learned lifting stage. The modality index is omitted.

Source selection. The inner weights in Eq. 6 come from a softmax over the reference logit zero and all valid local logits in Eq. 5. For outer attention, each source key $K_{j,\text{src}}^{(h)}$ projects $[G_j; E_j]$. A softmax over the null logit zero and source logits $\sqrt{d_h}\cos(Q_i^{(h)}, K_{j,\text{src}}^{(h)})$ gives $\alpha_i^{(h)}$. The null value is zero, and the received message is

$$C_i = W_{o,i}\,\text{Concat}_h\left[\sum_{j \neq i} \alpha_{ij}^{(h)} U_{ij}^{(h)}\right]. \tag{9}$$

Here $\alpha_{ij}^{(h)}$ is the weight on source j, Concat_h concatenates heads, and $W_{o,i}$ is a learned bias-free $d \to d$ output projection.

Matched-region sampling and risk. Training cycles through modalities and scales. Nominal region widths are $3, 7, 15$ for text and $15, 35, 75$ for audio/vision. A uniformly drawn center defines A, clipped to valid positions; a second center is drawn uniformly among candidates with the same valid width and the same positive count of newly erased positions. Their newly erased sets must differ but may overlap. If no candidate exists, only A contributes. Text erasure replaces tokens by `[UNK]`; audio/visual erasure zeros feature vectors. Text special tokens are excluded. A newly erased position previously contains a non-`[UNK]` token or a nonzero feature vector.

The polarity loss used both in the primary objective and in deletion risk is

$$\ell_{\text{pol}}(\ell, y) = w_0(y)\,\text{BCEWithLogits}(\ell, y_{\text{pol}}). \tag{10}$$

The binary target is $y_{\text{pol}} = \mathbf{1}\{y \geq 0\}$ for MOSI/MOSEI and $\mathbf{1}\{y > 0\}$ for CH-SIMS; $\mathbf{1}\{\cdot\}$ is the indicator function. The weight $w_0(y)$ is one except for neutral MOSEI targets, where it is 0.6. The current model measures original and additionally erased inputs in evaluation mode without gradients, reusing other modalities' encodings. The split is recomputed from detached original and erased features, so regional objectives update it with fixed targets and encoder inputs. Primary predictions and the unimodal objective retain end-to-end gradients.

Regional ranking. For newly erased positions $U^\star$ at the selected scale, noisy-OR gives $p_U = 1 - \prod_{\tau \in U^\star}(1 - q_{m,s,\tau})$ and $\rho_U = \mathrm{logit}(p_U)$. Pairwise differences are $\Delta D_{b,UV} = D_{b,U} - D_{b,V}$ and $\Delta\rho_{b,UV} = \rho_{b,U} - \rho_{b,V}$. In Eq. 3, let $w_{b,UV} = |\Delta D_{b,UV}|$ and $W = \sum_b \sum_{(U,V)\in\mathcal{P}_b} w_{b,UV}$. The normalized weight is $\overline{w}_{b,UV} = w_{b,UV}/\max(W, \epsilon)$, with numerical safeguard $\epsilon = 10^{-8}$. This pools all eligible $(A, 0)$, $(B, 0)$, and (A, B) comparisons in one weighted mean; an unmatched sample contributes only $(A, 0)$. Noisy-OR is evaluated in log-survival space, capping q below one by machine epsilon and flooring presence probabilities at the smallest positive normal value of the data type. No ranking margin or temperature is introduced.

Damage-weighted assignment. Let b index samples and $U \in \{A, B\}$ index eligible regions. For $X \in \{H_m, G_m^{\mathrm{seq}}\}$, let $\eta_X^{b,U}$ be the mean squared original–erased difference over all channels and originally valid positions; η_G abbreviates the response of G_m^{seq}. With $[z]_+ = \max(z, 0)$ and $w_{b,U} = [D_{b,U}]_+ \mathbf{1}\{\eta_H^{b,U} > \epsilon\}$, assignment minimizes

$$\mathcal{L}_{\mathrm{assign}} = \frac{\sum_{b,U} w_{b,U}\, \eta_G^{b,U} / \max(\eta_H^{b,U}, \epsilon)}{\max(\sum_{b,U} w_{b,U}, \epsilon)}. \tag{11}$$

The normalized weights in Eq. 4 are $\overline{\omega}_{b,U} = w_{b,U}/\max(\sum_{b',V} w_{b',V}, \epsilon)$. Empty valid support gives zero response energy. Ineligible regions receive zero weight, and the same safeguard ϵ is used throughout. Unlike the selected-scale noisy-OR scores, assignment measures the complete global stream and therefore involves all scales.

Prediction heads. The polarity logit in Section 3.4 is $\ell = \sum_m f_m(Z_m + C_m) + \Psi(G)$, where f_m is a modality-specific polarity head. The global interaction term uses Euclidean inner products $\langle\cdot,\cdot\rangle$:

$$\Psi(G) = \frac{1}{\sqrt{d}} \left[\sum_{m\in\{a,v\}} \langle A_m G_m, \tanh(W_c G_t + b_c)\rangle + \langle B_a G_a, B_v G_v\rangle \right]. \tag{12}$$

The $A_m, B_m \in \mathbb{R}^{d\times d}$ maps are learned bias-free projections; $W_c \in \mathbb{R}^{d\times d}$ and $b_c \in \mathbb{R}^d$ define the learned text-context map. The text head f_t is LayerNorm–MLP $d \to 64 \to 1$; audio/visual heads are bias-free linear maps. Intensity is $I = \mathrm{softplus}(f_I([G; E + C]))$. Its head flattens $[G_m; E_m + C_m]$ in text/audio/vision order and uses LayerNorm–MLP $6d \to d \to 1$. All MLPs use GELU. Since $2\sigma(\ell) - 1 = \tanh(\ell/2)$, the main-text readout matches the implementation. Scores are not clipped to the label range.

For the unimodal objective, an intensity head f_m^I with LayerNorm–MLP $d \to 64 \to 1$ gives $\widehat{y}_m = \mathrm{softplus}(f_m^I(Z_m))[2\sigma(f_m(Z_m)) - 1]$. With batch size B, $\mathcal{L}_{\mathrm{uni}} = (3B)^{-1} \sum_{b,m} (\widehat{y}_{b,m} - y_b)^2$. Each head uses the same multimodal target y_b. The no-exchange prediction uses the joint readout with $C = 0$ and receives only its squared-error term.

Loss coefficients. The effective coefficients in Eq. 7 are $\lambda_{\mathrm{aux}} = 0.2/7$ and $\lambda_I = 0.5$. The implementation divides the auxiliary sum by seven before applying its configured weight 0.2; the main-text coefficient absorbs this fixed scaling. Ranking, assignment, and unimodal prediction share the resulting coefficient.

Framework notation. Figure 2 omits indices where context is clear: in the split, $E_s = q_{m,s} \odot \delta_{m,s}$ is a per-scale sequence, distinct from summary residual E_m. Its z_0 denotes the zero-input BiGRU–LayerNorm response. Value symbols denote $V_{j,\mathrm{ref}}^{(h)} = W_{\mathrm{ref},j}^{(h)} E_j$ and $V_{j,s,\tau}^{(h)} = W_{\mathrm{loc},j}^{(h)}(S\delta_{j,s,\tau})$. Route is $\phi_{m,s}$; Proj denotes query/key maps. Circled C means concatenation; Scale multiplies cosine similarity by $\sqrt{d_h}$. LN, MLP, and BiGRU mean layer normalization, multilayer perceptron, and bidirectional gated recurrent unit. Prediction combines the displayed intensity branch with signed polarity (Section 3.4).

D.4 Baseline Implementations

Evaluation interfaces. We reproduce baseline models, including RECAP, SMCIR, and MIDAS, from their released implementations. To support fair comparisons across differing evaluation conventions, we keep each baseline's released architecture unchanged and standardize its evaluation

using the MIDAS protocol (Wen et al., 2026), including data splits, missing-input rules, and metric definitions. Each missing-rate curve and structured comparison evaluates a fixed, validation-selected prediction checkpoint. For transferred SMCIR/MOSI checkpoints, replay checks confirm identical inputs and masks, unchanged polarity decisions, and MAE differences below 10^{-4} against archived predictions. RECAP/MOSEI random-mask replays at $r = 0.5$ and $r = 0.9$ agree with saved native predictions from its newly selected checkpoints. Structured evaluations are also checked against native inference on the evaluation device.

MIDAS. The released model and loss are unchanged. Protocol adaptations comprise shared training masks and validation-based checkpoint selection; MOSEI uses per-batch learning-rate updates and the supplied 500-position audio length. Training runs for 20 epochs on MOSEI and 100 on MOSI/CH-SIMS. The MOSI curves and structured tests use checkpoints from fresh runs with the native MOSI loader and settings, with validation predictions retained for every epoch.

SMCIR. The public MOSI/MOSEI implementation is configured with aligned feature dimensions, disjoint optimizer groups, and batch-length handling in dynamic fusion. Training retains native L1-based stopping; shared validation MSE selects checkpoints. The released auxiliary-similarity flag remains zero. Because fusion depends on the batch, evaluation preserves batch size 32 and native sample order.

SMCIR on CH-SIMS. We extend the released MOSI/MOSEI implementation to CH-SIMS using Chinese BERT, text/audio/visual dimensions 768/33/709, and sequence lengths 39/400/55. Training/evaluation batches are 64/32 in native order, with FP32, Adam learning rate 2×10^{-5}, $\epsilon = 3 \times 10^{-8}$, weight decay 10^{-4}, gradient clipping at 2, and no learning-rate schedule. Shared training masks use seed 9001. Native rounded validation L1 controls stopping (patience 8, minimum improvement 10^{-6}); minimum clipped validation MAE at $r = 0.5$ after epoch 1 selects the checkpoint, with earliest ties. The same compatibility adaptations and zero auxiliary-similarity flag apply. Complete histories and validation/test predictions are retained for all three seeds. Random masks match deterministic replay of the shared-loader protocol; structured tests use DREAM's saved actual masks and sample identities.

RECAP on MOSEI. We evaluate the released 128-dimensional Transformer variant with 30–60 completion epochs and 30–100 prediction epochs per seed (patience 20 in both stages). Completion weights remain frozen during prediction training, and validation selects prediction checkpoints. All three runs are included.

D.5 Hardware and computational cost

Experiments use Python 3.9.18, PyTorch 2.2.2, CUDA 12.1, Transformers 4.30.0, and FP32. Training and diagnostics use driver-reported NVIDIA GeForce RTX 4090 GPUs (49,140 MiB) and RTX 4080 SUPER GPUs (32,760 MiB, approximately 32 GiB).

Each dataset's methods are timed on the same RTX 4080 SUPER instance with four CPU threads. FP32 timing disables TF32 and includes BERT with GPU-resident inputs: 10 warm-up calls precede 90 synchronized calls. Data loading and transfer are excluded.

Table 13: Inference cost at batch size 32. Parameters include all resident weights, including inactive training modules. Latency is the median of 90 calls; memory is peak GPU allocation.

	MOSI			MOSEI		CH-SIMS	
Metric	RECAP	Summary	**DREAM**	RECAP	**DREAM**	RECAP	**DREAM**
Params (M)↓	196.09	111.63	111.63	196.12	111.64	188.91	104.51
Latency (ms)↓	26.80	33.00	35.62	29.57	56.86	17.71	43.30
Memory (MiB)↓	1473	657	780	1491	823	1239	622

Full DREAM training on MOSEI takes 58.7–63.2 min across three seeds, including validation. Local reading adds 2.63 ms (8.0%) over Summary-only on MOSI. DREAM has lower allocated memory but longer batch latency than RECAP across datasets; all resident modules are counted.