FUSING EVIDENCE, LEARNING ACROSS VIEWS: MULTIMODAL SENTIMENT ANALYSIS WITH INCOMPLETE DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal sentiment analysis (MSA) is vulnerable to fine-grained missingness, where different corruption patterns retain different subsets of evidence from the same underlying sample. We propose Cross-view Evidence Learning and Alignment (CELA), a same-sample multi-corruption framework that aggregates evidence within each incomplete view and coordinates representations across independently corrupted views. CELA integrates modality-specific and common evidence through slot-based fusion, anchors incomplete-view representations to a task-supervised full-view representation, and explicitly promotes consistency among corrupted-view representations of the same utterance. At inference, CELA requires only a single incomplete observation, without explicitly reconstructing missing content. Experiments on MOSI, MOSEI, and SIMS show that CELA achieves the best missing-rate-averaged performance among the evaluated methods, with clear advantages under low-to-moderate missingness. These advantages also persist under fixed- and mixed-missing-rate training protocols, CIM-oriented nonidealities, and component-level physical ReRAM-based CIM execution.

1 INTRODUCTION

Multimodal sentiment analysis (MSA) estimates expressed sentiment from complementary Language, Visual, and Audio cues (Zadeh et al., 2016; Bagher Zadeh et al., 2018; Yu et al., 2020). In practical multimodal sensing pipelines, however, language tokens or audio/visual frames may become unavailable because of sensing, transmission, or synchronization failures. Token- and frame-level loss can remove useful information from these streams, motivating the study of multimodal sentiment analysis under incomplete observations (Yuan et al., 2021; Zhang et al., 2024). At a fixed missing rate, different corruption realizations retain different words, frames, and cross-modal correspondences. The sentiment label remains attached to the underlying utterance, but the evidence available for predicting it changes. Robust incomplete MSA therefore requires both effective use of the evidence retained within each observation and learning how task representations should relate across different observations of the same sample.

Existing methods address incomplete multimodal observations through feature reconstruction, dominant-modality correction, and uncertainty-aware representation or fusion mechanisms (Yuan et al., 2021; Zhang et al., 2024; Zhu et al., 2025). These approaches primarily specify how an individual incomplete observation should be processed. A complementary training question is how multiple independently corrupted observations of the same underlying sample should relate to one another. Applying task supervision independently to each corrupted observation does not explicitly exploit the fact that these observations originate from the same utterance. Same-sample corruptions provide an explicit relation for such coordination without requiring additional sentiment annotations: different views retain different subsets of evidence while preserving the same underlying task semantics. This relation is useful only to the extent that the retained evidence remains informative, since extreme missingness can make the original sentiment intrinsically ambiguous.

To exploit this relation, we propose Cross-view Evidence Learning and Alignment (CELA) for fine-grained incomplete MSA. Within each corrupted view, CELA retains modality-specific information, derives common evidence, and integrates the available evidence into a task representation.

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.

Across views, the full observation provides a training-time semantic reference, while independently corrupted incomplete realizations are directly coordinated at the representation level. CELA therefore couples within-view evidence learning and fusion with full-to-incomplete anchoring and incomplete-to-incomplete consistency. The full view and additional corrupted views are used only during training, whereas inference requires a single incomplete observation. Our contributions are threefold:

- We formulate same-sample multi-corruption learning for fine-grained incomplete MSA and instantiate it in CELA through within-view evidence learning, full-view anchoring, and cross-missing consistency.
- On MOSI, MOSEI, and SIMS, CELA achieves the best missing-rate-averaged overall performance among the evaluated methods, with consistent gains across the reported metrics and robust performance under additional missingness protocols.
- We further validate CELA under CIM-oriented nonidealities and physical ReRAM-based CIM execution, where its advantage remains observable under hardware nonidealities.

2 Related Work

2.1 Multimodal Representation Learning for Sentiment Analysis

Multimodal sentiment analysis has been widely studied through representation learning and cross-modal fusion. Early methods model higher-order interactions using tensor or low-rank fusion (Zadeh et al., 2017; Liu et al., 2018), recurrent architectures (Zadeh et al., 2018a;b), and factorized or translation-based representations (Tsai et al., 2019b; Pham et al., 2019). Attention-based methods model Language–nonverbal interaction through representation modulation, cross-modal attention, and pretrained Language models (Wang et al., 2019; Tsai et al., 2019a; Rahman et al., 2020). Other work improves representations by separating shared and modality-specific information (Hazarika et al., 2020), introducing auxiliary unimodal supervision or information-theoretic objectives (Yu et al., 2021; Han et al., 2021), and exploiting contrastive or consistency-based learning (Lin & Hu, 2022; Liu et al., 2022). Language-guided methods further leverage asymmetric modality contributions to refine nonverbal representations (Wang et al., 2023a;b; Zhang et al., 2023). Collectively, these studies provide effective mechanisms for learning multimodal evidence, but primarily model interactions among modalities within an observation rather than variations across different incomplete realizations of the same sample.

2.2 Multimodal Sentiment Analysis with Incomplete Data

Incomplete multimodal learning has been addressed through modality imagination, meta-learning, adaptive fusion, and modeling of imbalanced modality availability (Ma et al., 2021; Zhao et al., 2021; Ma et al., 2022; Sun et al., 2024a). In MSA, reconstruction-based methods recover missing information via feature reconstruction, missingness-aware encoding, or multi-level restoration (Yuan et al., 2021; Zeng et al., 2023; Sun et al., 2024b), while other methods improve incomplete representations through adversarial training, complete-to-incomplete alignment, self-distillation, knowledge transfer, and hierarchical representation learning (Yuan et al., 2024; Lin & Hu, 2023; Li et al., 2024a;b;c). Recent methods further exploit modality reliability: Language-dominated learning uses dense linguistic cues to guide corrupted representations (Zhang et al., 2024), whereas proxy-based learning models modality uncertainty in a probabilistic latent space (Zhu et al., 2025). More broadly, multi-view learning relates transformed views of the same instance through prediction, stop-gradient optimization, and representation regularization (Grill et al., 2020; Chen & He, 2021; Zbontar et al., 2021; Bardes et al., 2022); incomplete-MSA methods also connect complete and corrupted observations through representation attraction, alignment, or distillation (Sun et al., 2024b; Lin & Hu, 2023; Li et al., 2024c). These studies show the value of cross-observation transfer. Fine-grained stochastic corruption, however, naturally produces multiple incomplete views of the same utterance that share task semantics while retaining different evidence, motivating explicit coordination among incomplete observations alongside within-view evidence learning.

3 METHOD

Figure 1 summarizes the shared forward computation and training losses.

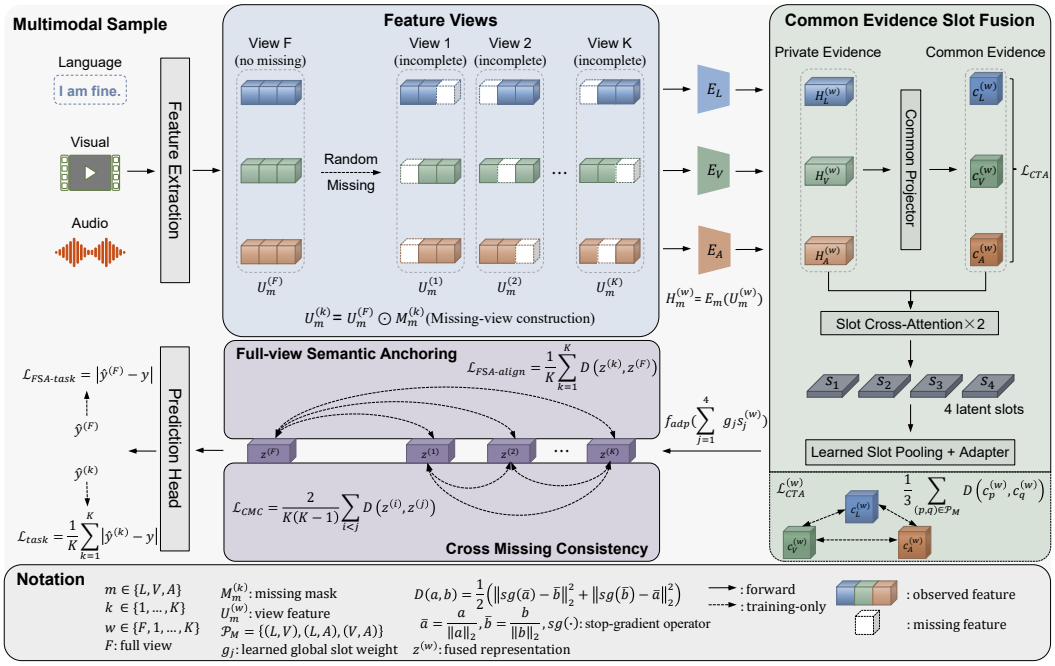


Figure 1: Overview of CELA. Shared encoders and CESF process a full view and K corrupted views during training, with FSA and CMC coordinating their representations. Inference requires only a single incomplete view.

3.1 PROBLEM FORMULATION

Let $U^{(F)} = \{U_L^{(F)}, U_V^{(F)}, U_A^{(F)}\}$ be the full preprocessed input of an utterance with sentiment target y. For each corrupted view $k \in \{1, \ldots, K\}$ and modality $m \in \{L, V, A\}$, a binary mask $M_m^{(k)}$ specifies retained positions (one) and removed positions (zero). We write $U_m^{(k)} = U_m^{(F)} \odot M_m^{(k)}$, where $\odot$ denotes mask application: removed Visual and Audio positions are zeroed and removed Language tokens become `[UNK]`, following the random-erasure convention of Zhang et al. (2024). Masks are sampled independently across modalities and views conditional on the missing rate. The same label supervises every view.

For $w \in \{F, 1, \ldots, K\}$, a modality-specific stack produces $H_m^{(w)} = E_m(U_m^{(w)}) \in \mathbb{R}^d$. Each E_m has its own parameters, reused across views. All subsequent fusion and prediction parameters are also shared across views. The full view is available during training only.

3.2 EVIDENCE LEARNING AND FUSION WITHIN A VIEW

Common Evidence Slot Fusion (CESF). CESF preserves $H_m^{(w)}$ as modality-specific evidence extracted before cross-modal interaction. To derive common evidence, it uses a coarse availability descriptor $\rho^{(w)} = [\chi_L^{(w)}, \chi_V^{(w)}, \chi_A^{(w)}, \bar{\chi}^{(w)}] \in \mathbb{R}^4$, where $\chi_m^{(w)}$ indicates modality availability and $\bar{\chi}^{(w)} = \frac{1}{3}\sum_m \chi_m^{(w)}$. This descriptor does not encode token/frame missing rates or positions. A projector P_c, shared across modalities and views, forms

$$c_m^{(w)} = P_c([H_m^{(w)}; \rho^{(w)}; e_m]) + r_m, \tag{1}$$

where $[\cdot;\cdot]$ denotes concatenation, $e_m \in \mathbb{R}^3$ is a fixed one-hot modality code, and $r_m \in \mathbb{R}^d$ is a learned role embedding. The six tokens are stacked as rows of an evidence bank:

$$\mathcal{E}^{(w)} = [H_L^{(w)}, H_V^{(w)}, H_A^{(w)}, c_L^{(w)}, c_V^{(w)}, c_A^{(w)}] \in \mathbb{R}^{6\times d}. \tag{2}$$

Keeping both token types allows fusion to access unimodal evidence without routing all information through the common pathway.

Four learned slots ($J = 4$) query this bank. For learned base states $b_j \in \mathbb{R}^d$ and an offset projector $P_{\text{off}} : \mathbb{R}^4 \to \mathbb{R}^{Jd}$, initialization is

$$s_j^{(w,0)} = b_j + [\text{reshape}_{J\times d}(P_{\text{off}}(\rho^{(w)}))]_j. \tag{3}$$

Two slot cross-attention blocks update the stacked states $S^{(w,0)}$:

$$S^{(w,\ell)} = \text{SCA}_\ell(S^{(w,\ell-1)}, \mathcal{E}^{(w)}), \qquad \ell = 1, 2. \tag{4}$$

Cross-attention uses the slots as queries and the bank as keys and values (Vaswani et al., 2017; Jaegle et al., 2021). Fine-grained corruption affects the evidence vectors; $\rho^{(w)}$ only conditions coarse availability. The final slots $s_j^{(w)} = [S^{(w,2)}]_j$ are pooled and transformed by an adapter:

$$z^{(w)} = f_{\text{adp}}\left(\sum_{j=1}^{J} g_j s_j^{(w)}\right), \qquad g_j = \frac{\exp(\alpha_j)}{\sum_{i=1}^{J}\exp(\alpha_i)}. \tag{5}$$

Here α_j are learned global logits; the pooling weights are shared across samples and views, while the slot contents depend on the input.

To regularize the projected common evidence, we introduce common-token agreement (CTA), which encourages cross-modal agreement among the common tokens within each view. For nonzero vectors, let $\bar{a} = a/\|a\|_2$ and $\bar{b} = b/\|b\|_2$, and define the symmetric stop-gradient distance

$$D(a, b) = \tfrac{1}{2}\big(\|\,\text{sg}(\bar{a}) - \bar{b}\|_2^2 + \|\,\text{sg}(\bar{b}) - \bar{a}\|_2^2\big), \tag{6}$$

where sg stops gradients through its argument. Both representations receive updates, one from each directional term. With modality pairs $\mathcal{P}_M = \{(L, V), (L, A), (V, A)\}$, the common-token agreement loss is

$$\mathcal{L}_{\text{CTA}} = \frac{1}{3(K+1)} \sum_{w\in\{F,1,\ldots,K\}} \sum_{(p,q)\in\mathcal{P}_M} D(c_p^{(w)}, c_q^{(w)}). \tag{7}$$

The per-view pair average is $\mathcal{L}_{\text{CTA}}^{(w)}$ in Figure 1; Eq. 7 averages it across views. The modality-specific evidence receives no direct agreement penalty.

3.3 Alignment across Corrupted Views

A shared head predicts $\hat{y}^{(w)} = f_{\text{pred}}(z^{(w)})$. Independent task supervision constrains the predictions; the following losses directly constrain the fused representations.

Full-view Semantic Anchoring (FSA). The full view supplies a supervised reference through $\mathcal{L}_{\text{FSA-task}} = \ell_{\text{task}}(\hat{y}^{(F)}, y)$, where ℓ_{task} denotes the dataset-specific regression loss defined in Section 3.4. Its representation is paired with every incomplete view:

$$\mathcal{L}_{\text{FSA-align}} = \frac{1}{K}\sum_{k=1}^{K} D\left(z^{(k)}, z^{(F)}\right). \tag{8}$$

Eq. 6 updates both full and incomplete representations.

Cross-Missing Consistency (CMC). CMC directly penalizes disagreement between incomplete representations of the same utterance:

$$\mathcal{L}_{\text{CMC}} = \frac{2}{K(K-1)} \sum_{1\le i<j\le K} D(z^{(i)}, z^{(j)}), \qquad K \ge 2. \tag{9}$$

For $K = 1$, $\mathcal{L}_{\text{CMC}} = 0$. FSA-align penalizes both the displacement of the corrupted-view centroid from the full-view representation and the dispersion among corrupted views, measured in the normalized representation space. CMC further penalizes this dispersion, allowing its strength to be adjusted without increasing the full-view anchoring penalty (Appendix A.2, Proposition 2).

3.4 Training Objective and Inference

We use L1 regression for MOSI and MOSEI and MSE for SIMS, denoted uniformly by $\ell_{\text{task}}(\hat{y}, y)$. The incomplete-view task loss is

$$\mathcal{L}_{\text{task}} = \frac{1}{K} \sum_{k=1}^{K} \ell_{\text{task}}\left(\hat{y}^{(k)}, y\right).$$

Training minimizes the sample-averaged objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_{\text{fs}}\mathcal{L}_{\text{FSA-task}} + \lambda_{\text{fa}}\mathcal{L}_{\text{FSA-align}} + \lambda_{\text{cmc}}\mathcal{L}_{\text{CMC}} + \lambda_{\text{cta}}\mathcal{L}_{\text{CTA}}. \tag{10}$$

The incomplete-view task loss has unit weight. In the formal configuration, we use $\lambda_{\text{fs}} = 0.5$, $\lambda_{\text{fa}} = 0.01$, $\lambda_{\text{cmc}} = 0.1$, and $\lambda_{\text{cta}} = 0.01$. The corresponding loss-weight analysis is provided in Appendix A.3. At inference, a single incomplete view passes through the shared modality encoders, CESF, and prediction head. No full observation, additional corrupted view, or reconstruction is required.

4 Experiments

4.1 Datasets

We evaluate on CMU-MOSI (Zadeh et al., 2016), CMU-MOSEI (Bagher Zadeh et al., 2018), and CH-SIMS (Yu et al., 2020), abbreviated MOSI, MOSEI, and SIMS. All contain Language, Visual, and Audio inputs. The train/validation/test splits used here contain 1,284/229/686, 16,326/1,871/4,659, and 1,368/456/457 utterances, respectively. Sentiment targets lie in $[-3, 3]$ for MOSI/MOSEI and $[-1, 1]$ for SIMS.

4.2 Evaluation Settings and Criteria

The main comparison uses ten matched missing rates, $\text{MR} \in \{0.0, 0.1, \ldots, 0.9\}$, with train MR equal to test MR. Random erasure follows the input convention in Section 3.1 (Zhang et al., 2024). CELA uses $K = 2$ unless specified otherwise. For each rate and seed (1111, 2222, 3333), we select the checkpoint with the highest validation BGSS and evaluate it on the test set. Main-table entries average over these three seeds and ten rates. Thus, these aggregates describe separately trained matched-rate models; Section 4.6 evaluates one checkpoint across test rates.

MOSI/MOSEI metrics are Acc-7, Acc-5, Acc-2, F1, MAE, and Pearson correlation (Corr); SIMS uses Acc-5 and Acc-3 in place of Acc-7 and Acc-5. For MOSI/MOSEI, binary results use Non0/Has0: Non0 excludes neutral targets, while Has0 retains them in the nonnegative class (Zhang et al., 2024). Classification and Corr are maximized; MAE is minimized. We define the Balanced Group Selection Score (BGSS) for this study to equally weight fine-grained classification, binary classification, and regression:

$$\text{BGSS} = \frac{1}{3}\left[\frac{A_1 + A_2}{2} + \frac{\text{Acc2} + \text{F1}}{2} + \frac{1}{2}\left(1 - \frac{\text{MAE}}{R} + \frac{\text{Corr} + 1}{2}\right)\right]. \tag{11}$$

Here (A_1, A_2) denotes (Acc-7, Acc-5) for MOSI/MOSEI and (Acc-5, Acc-3) for SIMS. Classification metrics enter as fractions; $R = 6$ for MOSI/MOSEI and $R = 2$ for SIMS. BGSS uses Non0 binary scores on MOSI/MOSEI. The constituent metrics are reported alongside this study-specific aggregate.

4.3 Robust Comparison

We retrain MISA (Hazarika et al., 2020), Self-MM (Yu et al., 2021), MMIM (Han et al., 2021), TETFN (Wang et al., 2023a), TFR-Net (Yuan et al., 2021), ALMT (Zhang et al., 2023), LNLN (Zhang et al., 2024), P-RMF (Zhu et al., 2025), and TF-Mamba (Li et al., 2025) under the evaluation in Section 4.2. Tables 1 and 2 report performance averaged over the evaluated missing rates, while Figure 2 resolves the results by missing rate. This comparison therefore evaluates both

Table 1: Robust comparison of the overall performance on MOSI and MOSEI. Acc-2 and F1 are reported as Non0/Has0. Lower MAE indicates better performance.

Method	MOSI							MOSEI						
	BGSS	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr	BGSS	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
MISA	0.5505	25.32	27.33	65.06/64.76	64.76/64.33	1.227	0.365	0.6429	43.68	44.18	71.55/72.99	68.11/70.64	0.743	0.412
Self-MM	0.5817	29.34	32.61	67.83/67.34	67.61/67.00	1.172	0.423	0.6776	46.47	47.26	74.89/74.27	73.97/74.14	0.691	0.509
MMIM	0.5752	28.22	31.23	67.10/66.61	66.66/66.05	1.171	0.429	0.6777	45.14	45.79	75.68/74.51	75.14/74.74	0.700	0.531
TETFN	0.5568	25.84	29.19	65.67/65.48	65.59/65.26	1.258	0.375	0.6669	44.26	45.15	74.37/73.21	73.59/73.31	0.729	0.498
TFR-Net	0.5653	27.43	30.66	66.93/65.91	64.80/63.64	1.217	0.393	0.6377	43.73	44.27	69.26/67.85	68.28/68.05	0.759	0.395
ALMT	0.5588	26.28	28.74	65.38/65.21	65.01/64.70	1.200	0.397	0.6683	44.93	45.52	74.25/72.99	73.34/72.99	0.713	0.497
LNLN	0.5647	27.11	29.79	66.03/65.65	65.11/64.61	1.191	0.413	0.6589	44.27	44.88	72.87/71.63	72.10/71.83	0.729	0.468
P-RMF	0.5679	27.23	29.62	66.99/66.50	66.83/66.22	1.221	0.408	0.6540	44.29	44.74	72.53/72.30	70.68/71.45	0.734	0.448
TF-Mamba	0.5733	27.90	30.86	67.17/66.58	67.20/66.50	1.196	0.416	0.6566	43.26	43.87	72.95/71.71	72.43/72.10	0.736	0.475
CELA	**0.6072**	**32.38**	**36.43**	**70.35/69.59**	**70.22/69.35**	**1.148**	**0.481**	**0.6926**	**47.04**	**48.00**	**77.10/76.42**	**76.36/76.35**	**0.676**	**0.567**

Table 2: Robust comparison of the overall performance on SIMS. Lower MAE indicates better performance.

Method	BGSS	Acc-5	Acc-3	Acc-2	F1	MAE↓	Corr
MISA	0.5828	26.51	55.84	70.44	62.30	0.562	0.254
Self-MM	0.6067	31.98	56.89	71.48	66.48	0.545	0.288
MMIM	0.6025	29.85	57.07	71.55	65.88	0.545	0.288
TETFN	0.5895	27.26	56.28	71.24	64.15	0.554	0.249
TFR-Net	0.5649	25.29	53.12	70.16	60.55	0.573	0.169
ALMT	0.6172	30.58	56.25	70.20	69.79	0.525	0.394
LNLN	0.6072	31.52	57.02	71.61	66.38	0.548	0.304
P-RMF	0.5949	30.08	56.20	70.56	63.93	0.559	0.282
TF-Mamba	0.6259	30.39	58.11	72.53	70.71	0.526	0.402
CELA	**0.6465**	**35.37**	**59.88**	**73.27**	**72.34**	**0.495**	**0.436**

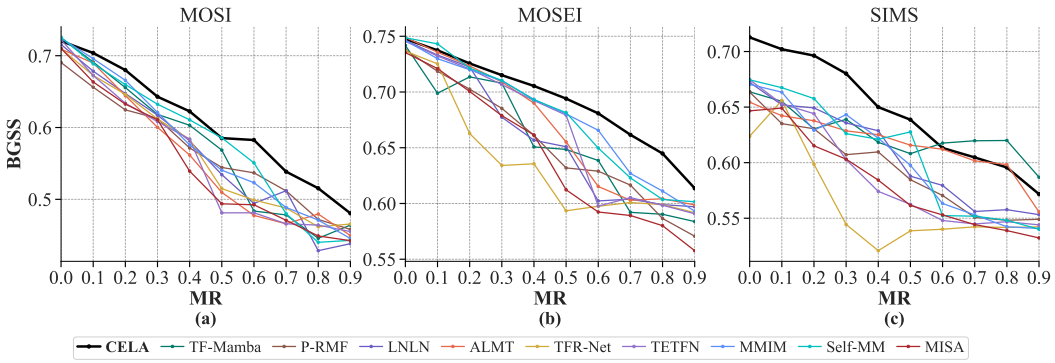


Figure 2: BGSS across matched missing rates on MOSI, MOSEI, and SIMS.

the overall robustness across missing conditions and how that robustness evolves as the retained multimodal evidence becomes increasingly sparse.

Among the compared methods, CELA achieves the highest macro-averaged BGSS on MOSI, MOSEI, and SIMS. The strongest competitor differs across datasets: CELA exceeds Self-MM by 0.0255 on MOSI, MMIM by 0.0149 on MOSEI, and TF-Mamba by 0.0206 on SIMS. The advantage is also reflected across all reported constituent metrics, including classification accuracy, F1, MAE, and Corr, rather than being driven by a single BGSS component. Notably, the best competing method changes from one dataset to another, while CELA remains the top method in the rate-averaged comparison across all three benchmarks, indicating that the gain is not tied to a particular baseline architecture or dataset. The per-rate curves in Figure 2 provide a finer-grained view of this advantage. CELA maintains a clear lead over much of the low-to-moderate missingness range, where sufficient but incomplete multimodal evidence remains available for cross-view learning and evidence fusion

to operate effectively. As the missing rate becomes severe, the gaps narrow and individual rankings can reverse, reflecting the increasing ambiguity of highly incomplete observations. Taken together, the tables and per-rate results show that CELA provides the strongest overall performance across the evaluated matched random-erasure conditions, with its main advantage concentrated in the practically relevant low-to-moderate missingness regime. Seed-level macro uncertainty and matched per-rate comparisons are reported in Appendix A.5.

4.4 Ablation Study

We conduct ablation studies on MOSI to assess the contributions of the main components and training objectives of CELA. Following the main evaluation protocol, Table 3 reports results macro-averaged over the ten matched missing rates. Removing CESF, CMC, and FSA-align reduces BGSS from 0.6072 to 0.5960, 0.5988, and 0.5920, respectively, while removing FSA entirely produces the largest drop to 0.5698.

Table 3: Ablation results on MOSI averaged over ten matched missing rates. Acc-2 and F1 are reported as Non0 / Has0. Lower MAE indicates better performance.

Method	BGSS	Acc-7	Acc-5	Acc-2	F1	MAE↓	Corr
CELA	**0.6072**	**32.38**	**36.43**	**70.35 / 69.59**	**70.22 / 69.35**	**1.148**	**0.481**
w/o CESF	0.5960	30.67	34.21	69.50 / 69.20	69.53 / 69.11	1.169	0.464
w/o CMC	0.5988	31.11	35.11	69.53 / 68.92	69.34 / 68.62	1.163	0.472
w/o FSA-align	0.5920	30.96	34.41	68.71 / 68.19	68.67 / 68.03	1.185	0.444
w/o FSA	0.5698	27.67	30.90	66.92 / 66.22	66.79 / 65.98	1.214	0.397

The ablations reveal a clear division of roles among the three components. CESF improves the organization and fusion of modality-specific and common evidence within each incomplete observation, whereas CMC provides an additional constraint directly between independently corrupted views. FSA contributes most strongly overall: the gap between w/o FSA-align and w/o FSA further shows that its full-view task supervision and representation alignment are complementary, with the former providing a task-discriminative reference and the latter constraining incomplete representations relative to that reference. The effects are not uniform across missing rates, suggesting that the components become useful under different levels of retained evidence rather than contributing as simple additive terms. Appendix A.6 provides the exact ablation configurations and rate-wise analysis, and further compares CELA with a near-parameter-matched Baseline under the same multi-view training exposure and task supervision to control for the generic benefit of multi-view training.

4.5 Effect of the Number of Corrupted Views

We examine the effect of the number of independently corrupted training views on MOSI by varying $K \in \{1, 2, 3, 4\}$ at three representative missing rates, $\mathrm{MR} \in \{0.1, 0.5, 0.9\}$. Each MR is trained and evaluated separately under the matched setting with three random seeds. For the same $(\mathrm{MR}, \mathrm{seed})$ pair, all values of K are evaluated using identical test masks for a controlled comparison. Table 4 reports BGSS together with training time and peak memory consumption.

Table 4: Effect of the number of corrupted views on MOSI. BGSS is reported as mean±std over three seeds. Avg. denotes the average over $\mathrm{MR} \in \{0.1, 0.5, 0.9\}$.

K	MR=0.1	MR=0.5	MR=0.9	Avg.	Time (s/ep.)	Mem. (MiB)
1	0.6573 ± 0.0107	0.5844 ± 0.0310	0.4701 ± 0.0101	0.5706 ± 0.0149	4.00	3982
2	$\mathbf{0.6958 \pm 0.0179}$	0.5890 ± 0.0084	0.4846 ± 0.0121	0.5898 ± 0.0062	5.71	5070
3	0.6928 ± 0.0103	$\mathbf{0.5916 \pm 0.0039}$	$\mathbf{0.4879 \pm 0.0138}$	$\mathbf{0.5908 \pm 0.0078}$	7.38	6169
4	0.6926 ± 0.0181	0.5841 ± 0.0108	0.4853 ± 0.0043	0.5873 ± 0.0081	11.94	7255

Increasing K from 1 to 2 improves the average BGSS from 0.5706 to 0.5898, showing a clear benefit from training with multiple corrupted realizations of the same sample. Increasing K further

to 3 yields only a marginal additional gain of 0.0010 in average BGSS: $K = 2$ remains better at MR=0.1, whereas $K = 3$ is slightly better at MR=0.5 and 0.9. This small gain comes with approximately 29% higher training time per epoch and an additional 1.10 GB of peak memory relative to $K = 2$. Increasing to $K = 4$ provides no further average improvement and raises training time to more than twice that of $K = 2$. We therefore adopt $K = 2$ as the default, as it captures most of the benefit of multi-view training while maintaining substantially lower training cost. Since $K = 1$ disables CMC, the gain from $K = 1$ to $K = 2$ reflects the overall effect of the multi-view training configuration rather than view count alone.

4.6 Robustness across Missing-Rate Training Protocols

The main matched-MR evaluation trains a separate model for each missing rate. We further evaluate CELA under two training protocols on MOSI, each using a single checkpoint per method and seed across $\mathrm{MR} \in \{0.0, 0.1, \ldots, 0.9\}$. The first is *fixed-MR training* at $\mathrm{MR} = 0.5$, a moderate missingness level motivated by real-world sensing observations: Xu et al. (2025) report mean and median missing fractions close to 50% in large-scale wearable sensor data. The second is *mixed-MR training*, following the dynamic random-missing paradigm used in LNLN and TF-Mamba. Each training sample is kept complete with probability 0.5; otherwise, its missing rate is sampled as $r \sim \mathcal{U}(0, 1)$. For CELA, corrupted views of the same sample share the sampled rate and use independently sampled masks.

For fixed-MR training, checkpoints are selected by validation BGSS at $\mathrm{MR} = 0.5$. For mixed-MR training, selection uses mean validation BGSS over a fixed bank covering nine missing rates, $\mathrm{MR} \in \{0.1, 0.2, \ldots, 0.9\}$. Each selected checkpoint is frozen before evaluation across all ten test missing rates.

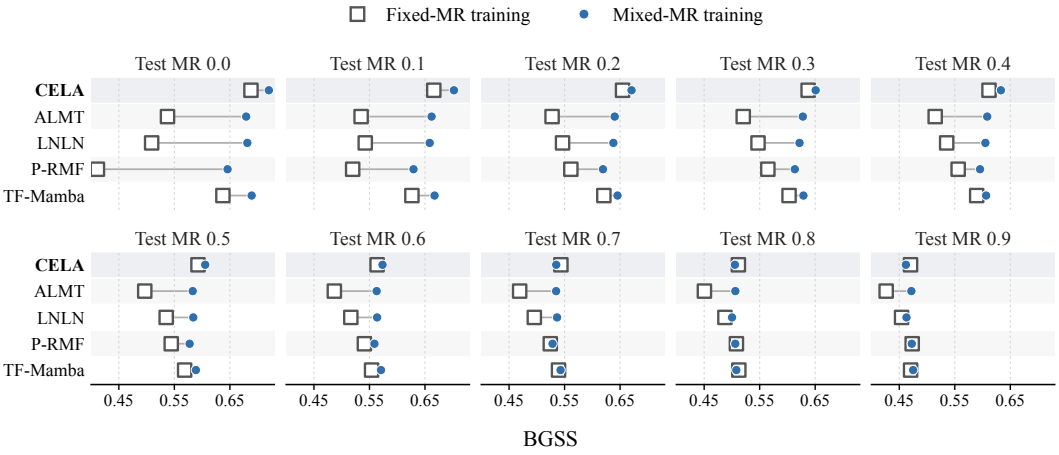
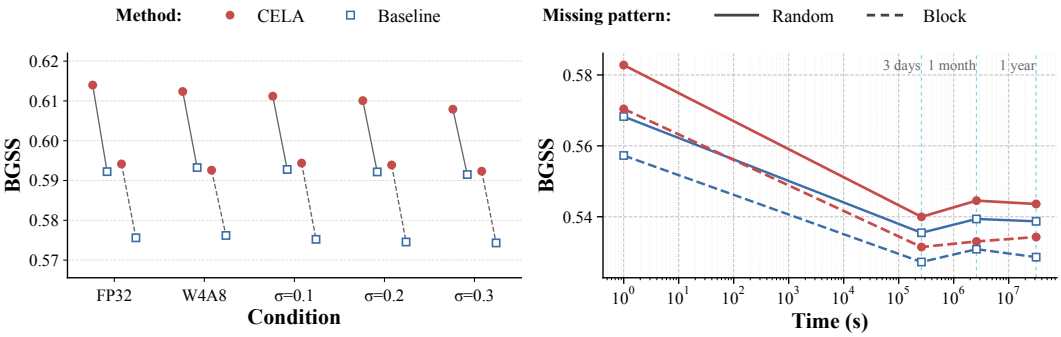


Figure 3: BGSS across test missing rates under fixed-MR ($\mathrm{MR} = 0.5$) and mixed-MR training on MOSI. Results are averaged over three seeds. Each method uses one validation-selected checkpoint per seed across all test rates.

We compare CELA with four selected baselines: ALMT, LNLN, P-RMF, and TF-Mamba. Figure 3 shows that CELA performs favorably across most test missing rates under fixed-MR training and remains among the strongest methods under mixed-MR training. These results extend the matched-MR comparison: CELA's advantage also holds when a single model is evaluated across varying missing rates, under both fixed-rate and mixed-rate training.

4.7 CIM-Oriented Simulation and Hardware Evaluation

We further evaluate CELA under CIM-oriented nonidealities on MOSI against a near-parameter-matched conventional Baseline. Both models are trained at MR$= 0.5$ using one complete and two independently corrupted views. Short-term simulation maps Fusion and Prediction with W4A8



(a) Short-term CIM robustness (b) Long-term CIM robustness

Figure 4: CIM-oriented simulation on MOSI at MR$= 0.5$ under Random and Block missingness. Short-term: Fusion+Prediction with W4A8 and weight perturbation. Long-term: prediction head with W4A8 and programming, drift, and read nonidealities.

Table 5: Physical ReRAM-based CIM prediction-head results on MOSI. Values are mean±sample standard deviation over three independent programming cycles. Lower MAE indicates better performance.

Method	Condition	BGSS	Acc-2	F1	MAE↓	Corr
CELA	Complete	**0.5783±0.0003**	**0.7581±0.0007**	**0.7595±0.0007**	1.2488±0.0041	**0.6598±0.0004**
Baseline	Complete	0.5705±0.0024	0.7342±0.0081	0.7353±0.0083	**1.2394±0.0027**	0.6163±0.0035
CELA	Block	**0.5324±0.0005**	**0.6824±0.0007**	**0.6841±0.0007**	**1.3072±0.0029**	**0.4635±0.0010**
Baseline	Block	0.5122±0.0021	0.6388±0.0041	0.6045±0.0067	1.3426±0.0031	0.4281±0.0027

quantization and multiplicative weight perturbations, while long-term simulation maps the prediction head with W4A8 quantization and models programming variation, conductance drift, and read noise. Random and Block masks use matched removal budgets. As shown in Figure 4, CELA maintains a positive mean BGSS margin over the Baseline across the evaluated short- and long-term conditions. Detailed mapping and nonideality settings are provided in Appendix A.9.

We additionally execute the $256 \to 256 \to 1$ prediction head on physical ReRAM-based CIM hardware using the same W4A8 path. All MOSI test samples are evaluated under Complete and Block (MR≈ 0.5) conditions over three independent programming cycles. Table 5 shows that under Block missingness CELA improves BGSS from 0.5122 to 0.5324, Acc-2 by 4.36 points, and F1 by 7.96 points over the Baseline, while also achieving lower MAE and higher Corr. Under the same prediction-head scope, HE-100 further reduces per-sample energy from 2.265 mJ on GPU (FP32) to approximately 0.995 mJ, corresponding to about a 56% reduction. Detailed latency, throughput, power, and energy measurements are reported in Appendix A.9. These results complement the simulation by showing that CELA's advantage remains observable under component-level physical CIM execution, together with the energy-efficiency potential of the mapped implementation.

Conclusion

We introduced CELA, a cross-view learning framework for incomplete multimodal sentiment analysis that combines within-view evidence fusion with full-view semantic anchoring and consistency across independently corrupted views, while requiring only a single incomplete observation at inference. Across MOSI, MOSEI, and SIMS, CELA achieves the strongest rate-averaged performance among the evaluated methods. Ablation studies support the complementary roles of evidence fusion, full-view supervision and alignment, and cross-missing consistency, while additional MOSI experiments show that the gains persist across missing-rate training protocols and the tested asymmetric and temporal missingness structures. CIM-oriented simulations and physical ReRAM prediction-head execution further preserve CELA's overall advantage under the evaluated nonidealities. Together, these results support learning across multiple corrupted views as an effective training principle for incomplete MSA without increasing inference-time view requirements.

AI USE STATEMENT

Generative AI tools were used to assist with literature search and related-work discovery, language polishing of the main paper, drafting selected portions of the appendix, and research execution, including organizing and formatting additional experimental results in the appendix and debugging research code. The authors retained responsibility for the problem formulation, method design, theoretical analysis, experimental protocols, implementation decisions, interpretation of results, and scientific conclusions. All reported results and manuscript content were reviewed and verified by the authors, who take full responsibility for the final paper and accompanying artifacts.

REPRODUCIBILITY STATEMENT

Appendices A.1–A.4 provide the architecture and implementation details, theoretical assumptions and proofs, loss-weight analysis, and complete experimental settings, including data processing, optimization, checkpoint selection, evaluation metrics, and baseline configurations. Appendices A.5–A.9 provide additional uncertainty analyses, ablation and robustness studies, and detailed CIM simulation and hardware protocols. Upon acceptance, we will publicly release the source code, training and evaluation scripts, and experimental configurations to facilitate reproducibility.

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 Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, June 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Chen_Exploring_Simple_Siamese_Representation_Learning_CVPR_2021_paper.html.

Zibo Chen, Ruxin Li, and Zilu Wang. Diversity-aware weight perturbation promotes robust adaptation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=5VoK9nUVSD.

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*, pp. 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1423.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9180–9192, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.723. URL https://aclanthology.org/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*, pp. 1122–1131, 2020. doi: 10.1145/3394171.3413678. URL https://doi.org/10.1145/3394171.3413678.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4651–4664. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html.

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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 10074–10082, 2024a. doi: 10.1609/aaai.v38i9.28871.

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*, volume 37, pp. 28515–28536, 2024b. doi: 10.52202/079017-0895.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12458–12468, 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 Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11252–11267, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.602. URL https://aclanthology.org/2025.findings-emnlp.602/.

Ronghao Lin and Haifeng Hu. Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 511–523. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-emnlp.36.

Ronghao Lin and Haifeng Hu. MissModal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023. doi: 10.1162/tacl_a_00628.

Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. Make acoustic and visual cues matter: CH-SIMS v2.0 dataset and AV-Mixup consistent module. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pp. 247–258, 2022. doi: 10.1145/3536221.3556630.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 2247–2256. Association for Computational Linguistics, 2018. doi: 10.18653/v1/P18-1209.

Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal learning with severely missing modality. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2302–2310, 2021. doi: 10.1609/aaai.v35i3.16330.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18177–18186, 2022.

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.

Jing Peng, Meiqi Yang, Qiong Zhang, and Xiaoxiao Li. S4m: S4 for multivariate time series forecasting with missing values. In *International Conference on Learning Representations*, 2025.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 6892–6899, 2019. doi: 10.1609/aaai.v33i01.33016892.

Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2359–2369. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.214.

Jun Sun, Xinxin Zhang, Shoukang Han, Yu-Ping Ruan, and Taihao Li. RedCore: Relative advantage aware cross-modal representation learning for missing modalities with imbalanced missing rates. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15173–15182, 2024a. doi: 10.1609/aaai.v38i13.29440.

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, 2024b. doi: 10.1109/TAFFC.2023.3274829.

Luan Tran, Xiaoming Liu, Jiayu Zhou, and Rong Jin. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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 *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569. Association for Computational Linguistics, 2019a. doi: 10.18653/v1/P19-1656.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. In *International Conference on Learning Representations*, 2019b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. arXiv:1706.03762, 2017. URL https://arxiv.org/abs/1706.03762.

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. URL https://doi.org/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.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7216–7223, 2019. doi: 10.1609/aaai.v33i01.33017216.

Maxwell A. Xu, Girish Narayanswamy, Kumar Ayush, Dimitris Spathis, Shun Liao, Shyam A. Tailor, Ahmed Metwally, A. Ali Heydari, Yuwei Zhang, Jake Garrison, Samy Abdel-Ghaffar, Xuhai Xu, Ken Gu, Jacob Sunshine, Ming-Zher Poh, Yun Liu, Tim Althoff, Shrikanth Narayanan, Pushmeet Kohli, Mark Malhotra, Shwetak Patel, Yuzhe Yang, James M. Rehg, Xin Liu, and

Daniel McDuff. LSM-2: Learning from incomplete wearable sensor data. *arXiv preprint arXiv:2506.05321*, 2025. URL https://arxiv.org/abs/2506.05321.

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 Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.343. URL https://aclanthology.org/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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 10790–10797, 2021. doi: 10.1609/aaai.v35i12.17289. URL https://ojs.aaai.org/index.php/AAAI/article/view/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. URL https://dl.acm.org/doi/10.1145/3474085.3475585.

Ziqi Yuan, Yihe Liu, Hua Xu, and Kai Gao. Noise imitation based adversarial training for robust multimodal sentiment analysis. *IEEE Transactions on Multimedia*, 26:529–539, 2024. doi: 10.1109/TMM.2023.3267882.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114. Association for Computational Linguistics, 2017. doi: 10.18653/v1/D17-1115.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018a. doi: 10.1609/aaai.v32i1.12021.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b. doi: 10.1609/aaai.v32i1.12024.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Robust multimodal sentiment analysis via tag encoding of uncertain missing modalities. *IEEE Transactions on Multimedia*, 25:6301–6314, 2023. doi: 10.1109/TMM.2022.3207572.

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 Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.49. URL https://aclanthology.org/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, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/65a39213d7d0e1eb5d192aa77e77eeb7-Abstract-Conference.html.

Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 2608–2618. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.203.

Aoqiang Zhu, Min Hu, Xiaohua Wang, Jiaoyun Yang, Yiming Tang, and Ning An. Proxy-driven robust multimodal sentiment analysis with incomplete data. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22123–22138, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1075. URL https://aclanthology.org/2025.acl-long.1075/.

A Additional Details and Results

A.1 Implementation Details of CELA

Table 6 summarizes the architecture used in the formal CELA experiments. All full and corrupted views share the same model parameters, the multimodal representation dimension is $d = 256$, and the complete model contains 113,819,141 parameters.

Table 6: Key architectural settings of CELA.

Component	Setting
Representation dimension	$d = 256$
Modality encoder	Mean pooling + Linear + residual FFN + one-layer BiGRU
Common projector	$263 \to 256 \to 256$
Latent slots	$J = 4$
Slot cross-attention	2 blocks, 1 head
Slot pooling	Learned global softmax weighting
Adapter	Rank 32, scale 0.1
Prediction head	$256 \to 256 \to 1$

Each modality sequence is first aggregated by mean pooling over valid sequence positions. The resulting modality-level representation is then processed by the linear projection, residual feed-forward block, and one-layer bidirectional GRU before multimodal fusion. The availability descriptor provides only coarse modality availability and does not encode missing rates or missing positions; all full and corrupted views share the same model parameters.

A.2 Theoretical Analysis

For nonzero representations, let $u_k = z^{(k)}/\|z^{(k)}\|_2$, $v = z^{(F)}/\|z^{(F)}\|_2$, $\bar{u} = K^{-1}\sum_k u_k$, and $V_K = K^{-1}\sum_k \|u_k - \bar{u}\|_2^2$. Write $A_K = \mathcal{L}_{\text{FSA-align}}$ and $C_K = \mathcal{L}_{\text{CMC}}$. The symmetric distance $D_{\text{sg}} = \frac{1}{2}(\|\operatorname{sg}(u) - v\|_2^2 + \|\operatorname{sg}(v) - u\|_2^2)$ has value $\|u - v\|_2^2$; its gradient is treated below. Finite-view identities hold for each realization; population statements fix model parameters and a deterministic forward map, conditioning on utterance x and missing rate q.

Corruption-induced dispersion.

Proposition 1 (The statistic constrained by CMC)**.** *For $K \geq 2$, with $S_K^2 = (K-1)^{-1}\sum_k \|u_k - \bar{u}\|_2^2$,*

$$C_K = \frac{2K}{K-1} V_K = 2S_K^2. \tag{12}$$

If $U_1, \ldots, U_K$ are conditionally i.i.d given (x, q), with conditional mean μ and covariance Σ, then

$$\mathbb{E}[C_K \mid x, q] = 2\operatorname{tr}\Sigma, \qquad \mathbb{E}[V_K \mid x, q] = \frac{K-1}{K}\operatorname{tr}\Sigma. \tag{13}$$

Proof. Expanding both sides gives $\sum_{i<j}\|u_i - u_j\|_2^2 = K\sum_i \|u_i\|_2^2 - \|\sum_i u_i\|_2^2 = K\sum_i \|u_i - \bar{u}\|_2^2$, which proves Eq. 12. Conditional independence gives $\mathbb{E}[\|U_i - U_j\|_2^2 \mid x, q] = 2(\mathbb{E}[\|U\|_2^2 \mid x, q] - \|\mu\|_2^2) = 2\operatorname{tr}\Sigma$. Averaging the pairs and using equation 12 proves equation 13. □

Thus, CMC is exactly proportional to the empirical dispersion of normalized corrupted-view representations and estimates conditional corruption-induced dispersion under the stated sampling assumption.

Joint geometry of FSA and CMC.

Proposition 2 (Anchoring and dispersion reweighting)**.** *For $K \geq 1$, and then for $K \geq 2$ with fixed nonnegative weights a, b,*

$$A_K = \|\bar{u} - v\|_2^2 + V_K, \tag{14}$$

$$G := aA_K + bC_K = a\|\bar{u} - v\|_2^2 + \left(a + \frac{2Kb}{K-1}\right) V_K. \tag{15}$$

At differentiable points, backpropagation through the original detached losses and ordinary differentiation of G satisfy, for shared parameters θ,

$$\nabla_\theta^{\mathrm{sg}}(aA_K + bC_K) = \tfrac{1}{2}\nabla_\theta G. \tag{16}$$

Proof. Expand $u_k - v = (u_k - \bar{u}) + (\bar{u} - v)$ and average squared norms. The cross term vanishes since $\sum_k (u_k - \bar{u}) = 0$, proving Eq. 14; substitute Eq. 12 to obtain Eq. 15. For one edge (u, v), the detached partial derivatives are $u - v$ and $v - u$, half those of $\|u - v\|_2^2$. With shared-parameter Jacobians J_u, J_v, backpropagation is $J_u^\top(u - v) + J_v^\top(v - u)$, half the ordinary gradient. Fixed weighted summation proves Eq. 16, including gradients through the trainable full-view branch. □

Hence, FSA-align jointly controls displacement from the full-view anchor and dispersion among corrupted views, while CMC further increases the dispersion penalty. The stop-gradient form preserves the same forward distance while scaling its shared-parameter first-order gradient by one half.

Conditional prediction stability.

Proposition 3 (Stability with radial control)**.** *Let the shared evaluation head h be L-Lipschitz on the relevant representations, $r_k = \|z^{(k)}\|_2$, $r_F = \|z^{(F)}\|_2$, and*

$$R_K = \frac{1}{K}\sum_k (r_k - r_F)^2, \qquad T_K = R_K + \frac{1}{K}\sum_k r_k r_F \|u_k - v\|_2^2. \tag{17}$$

Then $K^{-1}\sum_k |h(z^{(k)}) - h(z^{(F)})|^2 \leq L^2 T_K$; if all radii are at most B, then $T_K \leq R_K + B^2 A_K$. For the same target y and $e_F = |h(z^{(F)}) - y|$,

$$\frac{1}{K}\sum_k |h(z^{(k)}) - y| \leq e_F + L\sqrt{T_K}, \tag{18}$$

$$\frac{1}{K}\sum_k (h(z^{(k)}) - y)^2 \leq (e_F + L\sqrt{T_K})^2. \tag{19}$$

Proof. Unit norms give $\|z^{(k)} - z^{(F)}\|_2^2 = (r_k - r_F)^2 + r_k r_F \|u_k - v\|_2^2$. Apply Lipschitz continuity and average; the bound using B follows immediately. The triangle inequality and Cauchy–Schwarz give Eq. 18. The triangle inequality for the normalized Euclidean norm of K residuals gives $\sqrt{K^{-1}\sum_k (h(z^{(k)}) - y)^2} \leq e_F + L\sqrt{T_K}$; squaring proves Eq. 19. □

Equations 18–19 apply to the MOSI/MOSEI L1 and SIMS MSE task losses, respectively, while retaining the full-view error. Because the head consumes unnormalized z, normalized alignment alone does not guarantee prediction equality; the bound also depends on radial variation.

These results characterize the geometry of the training objective and conditional prediction stability. They do not imply unconditional generalization, recovery of erased information, monotonic improvement with K, or robustness to hardware perturbations.

A.3 Loss-Weight Analysis

We vary one auxiliary coefficient at a time on MOSI while retaining the other three at the formal values in Section 3.4. The formal coefficients were fixed before this analysis, and the sweep is used to assess their sensitivity over a broad range of alternatives. Each setting is evaluated over the ten matched missing rates and three seeds using validation-BGSS-selected checkpoints; Figure 5 reports the resulting rate- and seed-averaged Test BGSS. The same coefficients are used for MOSEI and SIMS without dataset-specific tuning. The adopted coefficients are marked explicitly; these sweeps compare the tested alternatives and do not establish a global optimum.

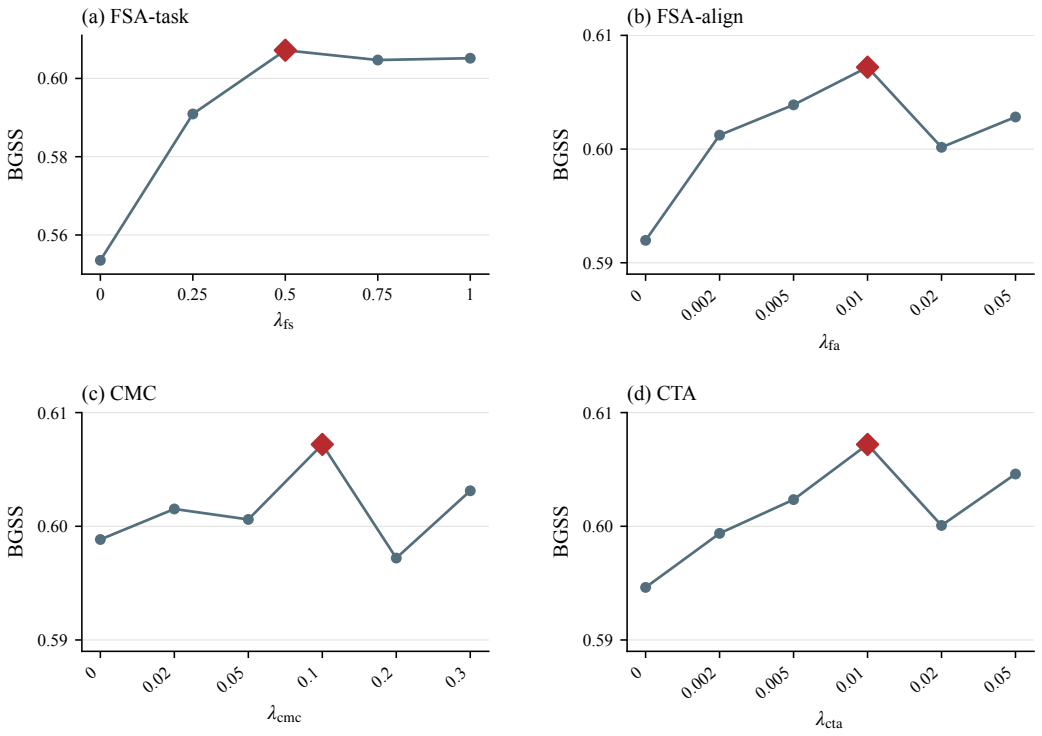


Figure 5: Loss-weight sweeps on MOSI. The marked point in each panel denotes the adopted coefficient.

The adopted values also achieve the highest tested BGSS in their respective one-at-a-time sweeps. In particular, the broader CMC sweep from $\lambda_{\text{cmc}} = 0$ to 0.3 continues to favor the adopted $\lambda_{\text{cmc}} = 0.1$, despite the non-monotonic response at larger weights. The results therefore support the adopted configuration over the tested coefficient ranges while also showing that the auxiliary objectives have different sensitivity profiles.

A.4 Experimental Setup and Reproducibility Details

Unless otherwise stated, CELA uses $K = 2$ corrupted views and seeds 1111, 2222, and 3333. This section summarizes the data processing, optimization, evaluation, and baseline configurations used in the experiments.

Data and corruption. Table 7 summarizes the processed inputs. Language uses BERT token inputs (Devlin et al., 2019), with `bert-base-uncased` for MOSI/MOSEI and `bert-base-chinese` for SIMS; Audio and Visual inputs are pre-extracted sequences. We use the pre-extracted Audio and Visual features distributed with the processed benchmark datasets.

Table 7: Dataset splits and processed input configurations used in the formal experiments. Language reports the BERT input shape, while Visual and Audio report sequence length $\times$ feature dimension.

Dataset	Train / Valid / Test	Language	Visual	Audio
MOSI	1284 / 229 / 686	3×50	500×20	375×5
MOSEI	16326 / 1871 / 4659	3×50	500×35	500×74
SIMS	1368 / 456 / 457	3×39	55×709	400×33

Eligible Language tokens are replaced by `[UNK]`, excluding special tokens and padding. Removed Audio and Visual frames are zeroed. Masks are sampled independently across modalities and corrupted views conditional on the missing rate.

Optimization. Table 8 summarizes the formal CELA training configuration. Loss coefficients are given in Section 3.4.

Table 8: Training hyperparameters of CELA on the three datasets.

	MOSI	MOSEI	SIMS
Optimizer	AdamW	AdamW	AdamW
Epochs	100	100	100
Batch Size	32	32	32
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
BERT Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-5}
Weight Decay	1×10^{-4}	1×10^{-4}	1×10^{-4}
Gradient Clipping	1.0	1.0	1.0
Task Loss	L1	L1	MSE
Corrupted Views K	2	2	2
Seed	1111, 2222, 3333	1111, 2222, 3333	1111, 2222, 3333

All CELA experiments are conducted on an NVIDIA L40 GPU using Python 3.8.20, PyTorch 1.12.1, and CUDA 11.3 in FP32 precision. The loss coefficients are given in Section 3.4.

Checkpoint selection. For the main matched-MR evaluation, one model is trained per missing rate and seed, and the checkpoint with the highest validation BGSS is evaluated at the corresponding test missing rate. Protocol-specific selection rules for auxiliary studies are described in their respective subsections.

Metric implementation. We use a single canonical metric implementation for the reported evaluations. Table 9 records the exact classification and regression conventions used by this implementation.

Table 9: Evaluation conventions.

Metric	Convention		
	MOSI	MOSEI	SIMS
Acc-7	Clip $[-3, 3]$ + round	Clip $[-3, 3]$ + round	–
Acc-5	Clip $[-2, 2]$ + round	Clip $[-2, 2]$ + round	Fixed bins
Acc-3	–	–	Fixed bins
Acc-2	Non0 / Has0	Non0 / Has0	Fixed bins
F1	Weighted, Non0 / Has0	Weighted, Non0 / Has0	Weighted
MAE / Corr	Raw values	Raw values	Raw values
BGSS	Eq. 11, Non0	Eq. 11, Non0	Eq. 11

For MOSI/MOSEI, Non0 excludes $y = 0$ with $y > 0$ as positive, whereas Has0 retains $y = 0$ with $y \geq 0$ as positive. SIMS uses fixed bins after clipping to $[-1, 1]$; MOSI/MOSEI use NumPy ties-to-even rounding. BGSS follows Eq. 11 and uses Non0 binary scores for MOSI/MOSEI.

Baseline implementations. Table 10 summarizes the implementation source and retained model-specific settings for each baseline.

Table 10: Baseline implementations and key model-specific settings.

Method	Implementation	Key settings
MISA	Open-source MMSA implementation	Hidden size 128; dropout 0.2; auxiliary-loss weights $0.8/0.1/0.3/1/1$
Self-MM	Open-source MMSA implementation	Audio/Visual LSTM sizes 32/64; fusion hidden size 128
MMIM	Open-source MMSA implementation	Audio/Visual recurrent size 32; projection size 128; NCE/LLD weights $0.1/0.1$
TETFN	Open-source MMSA implementation	Destination dimension 30; 6 heads; fusion hidden size 128; auxiliary weight 1
TFR-Net	Open-source MMSA implementation	3 fusion GRU layers; classifier size 128; generator weights $5/2/20$; similarity weight 5
ALMT	Open-source MMSA implementation	Feature size 128; AHL depth 3; fusion depth 2; 8 heads
LNLN	Official open-source implementation	Hidden size 128; depth 2; 8 heads; loss weights $0.9/0.8/0.1/1$
P-RMF	Paper-based reimplementation	Hidden size 128; 8 proxy tokens; 4 PDDI layers; 8 heads; loss weights $1/0.5/0.1$
TF-Mamba	Official open-source implementation	$d_{\text{state}} = 12/16/16$, expansion $= 4/4/2$, $\alpha = 0.7/0.3/1.0$; bidirectional; convolution width 4

All baselines use the same processed data splits, corruption protocol, random seeds, validation-based checkpoint selection, and evaluation metrics. We retain the model-specific configurations above without additional hyperparameter search.

A.5 Additional Main-Comparison Results

Metrics are averaged over Test MR$= 0.0, \ldots, 0.9$ for each seed and reported as mean $\pm$ sample standard deviation across three seeds. Bold denotes the best mean result and underline denotes the strongest baseline.

Table 11: Three-seed macro uncertainty on MOSI. Acc-2 and F1 report Non0 and Has0 separately; lower MAE is better.

Method	BGSS	Acc-7	Acc-5	Acc-2 Non0	Acc-2 Has0
MISA	0.5505 ± 0.0075	25.32 ± 0.29	27.33 ± 0.73	65.06 ± 0.97	64.76 ± 0.84
Self-MM	$\underline{0.5817 \pm 0.0066}$	$\underline{29.34 \pm 0.56}$	$\underline{32.61 \pm 0.86}$	$\underline{67.83 \pm 0.60}$	$\underline{67.34 \pm 0.66}$
MMIM	0.5752 ± 0.0180	28.22 ± 1.58	31.23 ± 2.00	67.10 ± 2.34	66.61 ± 1.93
TETFN	0.5568 ± 0.0030	25.84 ± 0.46	29.19 ± 0.50	65.67 ± 0.47	65.48 ± 0.40
TFR-Net	0.5653 ± 0.0179	27.43 ± 0.74	30.66 ± 1.42	66.93 ± 3.18	65.91 ± 2.49
ALMT	0.5588 ± 0.0039	26.28 ± 0.72	28.74 ± 0.70	65.38 ± 0.74	65.21 ± 0.71
LNLN	0.5647 ± 0.0116	27.11 ± 0.86	29.79 ± 1.32	66.03 ± 1.66	65.65 ± 1.58
P-RMF	0.5679 ± 0.0122	27.23 ± 1.47	29.62 ± 1.87	66.99 ± 1.10	66.50 ± 1.15
TF-Mamba	0.5733 ± 0.0106	27.90 ± 1.18	30.86 ± 1.38	67.17 ± 1.25	66.58 ± 0.95
CELA	$\mathbf{0.6072 \pm 0.0082}$	$\mathbf{32.38 \pm 0.85}$	$\mathbf{36.43 \pm 0.81}$	$\mathbf{70.35 \pm 1.25}$	$\mathbf{69.59 \pm 1.02}$

Method	F1 Non0	F1 Has0	MAE↓	Corr
MISA	64.76 ± 1.17	64.33 ± 1.05	1.227 ± 0.012	0.365 ± 0.025
Self-MM	$\underline{67.61 \pm 0.63}$	$\underline{67.00 \pm 0.67}$	1.172 ± 0.012	0.423 ± 0.032
MMIM	66.66 ± 2.72	66.05 ± 2.33	$\underline{1.171 \pm 0.032}$	$\underline{0.429 \pm 0.033}$
TETFN	65.59 ± 0.65	65.26 ± 0.58	1.258 ± 0.004	0.375 ± 0.013
TFR-Net	64.80 ± 3.45	63.64 ± 2.81	1.217 ± 0.035	0.393 ± 0.030
ALMT	65.01 ± 0.85	64.70 ± 0.85	1.200 ± 0.009	0.397 ± 0.022
LNLN	65.11 ± 1.94	64.61 ± 1.88	1.191 ± 0.018	0.413 ± 0.032
P-RMF	66.83 ± 1.17	66.22 ± 1.25	1.221 ± 0.028	0.408 ± 0.029
TF-Mamba	67.20 ± 1.21	66.50 ± 0.92	1.196 ± 0.032	0.416 ± 0.024
CELA	$\mathbf{70.22 \pm 1.44}$	$\mathbf{69.35 \pm 1.22}$	$\mathbf{1.148 \pm 0.025}$	$\mathbf{0.481 \pm 0.006}$

Table 12: Three-seed macro uncertainty on MOSEI. Acc-2 and F1 report Non0 and Has0 separately; lower MAE is better.

Method	BGSS	Acc-7	Acc-5	Acc-2 Non0	Acc-2 Has0
MISA	0.6429 ± 0.0077	43.68 ± 0.32	44.18 ± 0.22	71.55 ± 0.91	72.99 ± 0.61
Self-MM	0.6776 ± 0.0077	46.47 ± 0.60	47.26 ± 0.67	74.89 ± 0.85	74.27 ± 1.22
MMIM	0.6777 ± 0.0058	45.14 ± 0.35	45.79 ± 0.33	75.68 ± 0.66	74.51 ± 0.56
TETFN	0.6669 ± 0.0021	44.26 ± 0.46	45.15 ± 0.41	74.37 ± 0.70	73.21 ± 1.33
TFR-Net	0.6377 ± 0.0064	43.73 ± 0.42	44.27 ± 0.53	69.26 ± 0.85	67.85 ± 1.65
ALMT	0.6683 ± 0.0047	44.93 ± 0.11	45.52 ± 0.24	74.25 ± 0.91	72.99 ± 0.61
LNLN	0.6589 ± 0.0172	44.27 ± 1.36	44.88 ± 1.44	72.87 ± 2.21	71.63 ± 2.16
P-RMF	0.6540 ± 0.0009	44.29 ± 0.28	44.74 ± 0.32	72.53 ± 0.10	72.30 ± 0.88
TF-Mamba	0.6566 ± 0.0096	43.26 ± 0.69	43.87 ± 0.95	72.95 ± 1.20	71.71 ± 0.61
CELA	$\mathbf{0.6926 \pm 0.0019}$	$\mathbf{47.04 \pm 0.22}$	$\mathbf{48.00 \pm 0.15}$	$\mathbf{77.10 \pm 0.36}$	$\mathbf{76.42 \pm 0.58}$

Method	F1 Non0	F1 Has0	MAE↓	Corr
MISA	68.11 ± 1.86	70.64 ± 0.92	0.743 ± 0.010	0.412 ± 0.033
Self-MM	73.97 ± 0.93	74.14 ± 1.18	0.691 ± 0.011	0.509 ± 0.028
MMIM	75.14 ± 0.83	74.74 ± 0.64	0.700 ± 0.011	0.531 ± 0.024
TETFN	73.59 ± 0.48	73.31 ± 1.04	0.729 ± 0.006	0.498 ± 0.013
TFR-Net	68.28 ± 1.40	68.05 ± 1.21	0.759 ± 0.010	0.395 ± 0.024
ALMT	73.34 ± 1.18	72.99 ± 0.70	0.713 ± 0.006	0.497 ± 0.018
LNLN	72.10 ± 2.26	71.83 ± 2.05	0.729 ± 0.028	0.468 ± 0.056
P-RMF	70.68 ± 0.27	71.45 ± 0.68	0.734 ± 0.002	0.448 ± 0.016
TF-Mamba	72.43 ± 1.34	72.10 ± 0.71	0.736 ± 0.015	0.475 ± 0.032
CELA	$\mathbf{76.36 \pm 0.29}$	$\mathbf{76.35 \pm 0.49}$	$\mathbf{0.676 \pm 0.005}$	$\mathbf{0.567 \pm 0.001}$

Table 13: Three-seed macro uncertainty on SIMS. Lower MAE is better.

Method	BGSS	Acc-5	Acc-3	Acc-2
MISA	0.5828 ± 0.0103	26.51 ± 2.11	55.84 ± 0.23	70.44 ± 0.33
Self-MM	0.6067 ± 0.0013	31.98 ± 0.59	56.89 ± 0.65	71.48 ± 0.31
MMIM	0.6025 ± 0.0079	29.85 ± 0.92	57.07 ± 0.65	71.55 ± 0.50
TETFN	0.5895 ± 0.0054	27.26 ± 1.07	56.28 ± 0.55	71.24 ± 0.37
TFR-Net	0.5649 ± 0.0006	25.29 ± 0.37	53.12 ± 2.05	70.16 ± 0.05
ALMT	0.6172 ± 0.0322	30.58 ± 2.46	56.25 ± 4.37	70.20 ± 5.43
LNLN	0.6072 ± 0.0143	31.52 ± 0.16	57.02 ± 1.11	71.61 ± 1.29
P-RMF	0.5949 ± 0.0035	30.08 ± 1.08	56.20 ± 0.39	70.56 ± 0.20
TF-Mamba	0.6259 ± 0.0054	30.39 ± 0.33	58.11 ± 0.97	72.53 ± 0.21
CELA	$\mathbf{0.6465 \pm 0.0042}$	$\mathbf{35.37 \pm 1.03}$	$\mathbf{59.88 \pm 0.60}$	$\mathbf{73.27 \pm 0.22}$

Method	F1	MAE↓	Corr
MISA	62.30 ± 1.98	0.562 ± 0.010	0.254 ± 0.024
Self-MM	66.48 ± 1.09	0.545 ± 0.004	0.288 ± 0.016
MMIM	65.88 ± 1.19	0.545 ± 0.009	0.288 ± 0.027
TETFN	64.15 ± 0.36	0.554 ± 0.010	0.249 ± 0.043
TFR-Net	60.55 ± 0.65	0.573 ± 0.004	0.169 ± 0.018
ALMT	69.79 ± 4.37	0.525 ± 0.029	0.394 ± 0.047
LNLN	66.38 ± 2.47	0.548 ± 0.020	0.304 ± 0.055
P-RMF	63.93 ± 1.14	0.559 ± 0.005	0.282 ± 0.007
TF-Mamba	70.71 ± 0.47	0.526 ± 0.012	0.402 ± 0.017
CELA	$\mathbf{72.34 \pm 0.47}$	$\mathbf{0.495 \pm 0.006}$	$\mathbf{0.436 \pm 0.008}$

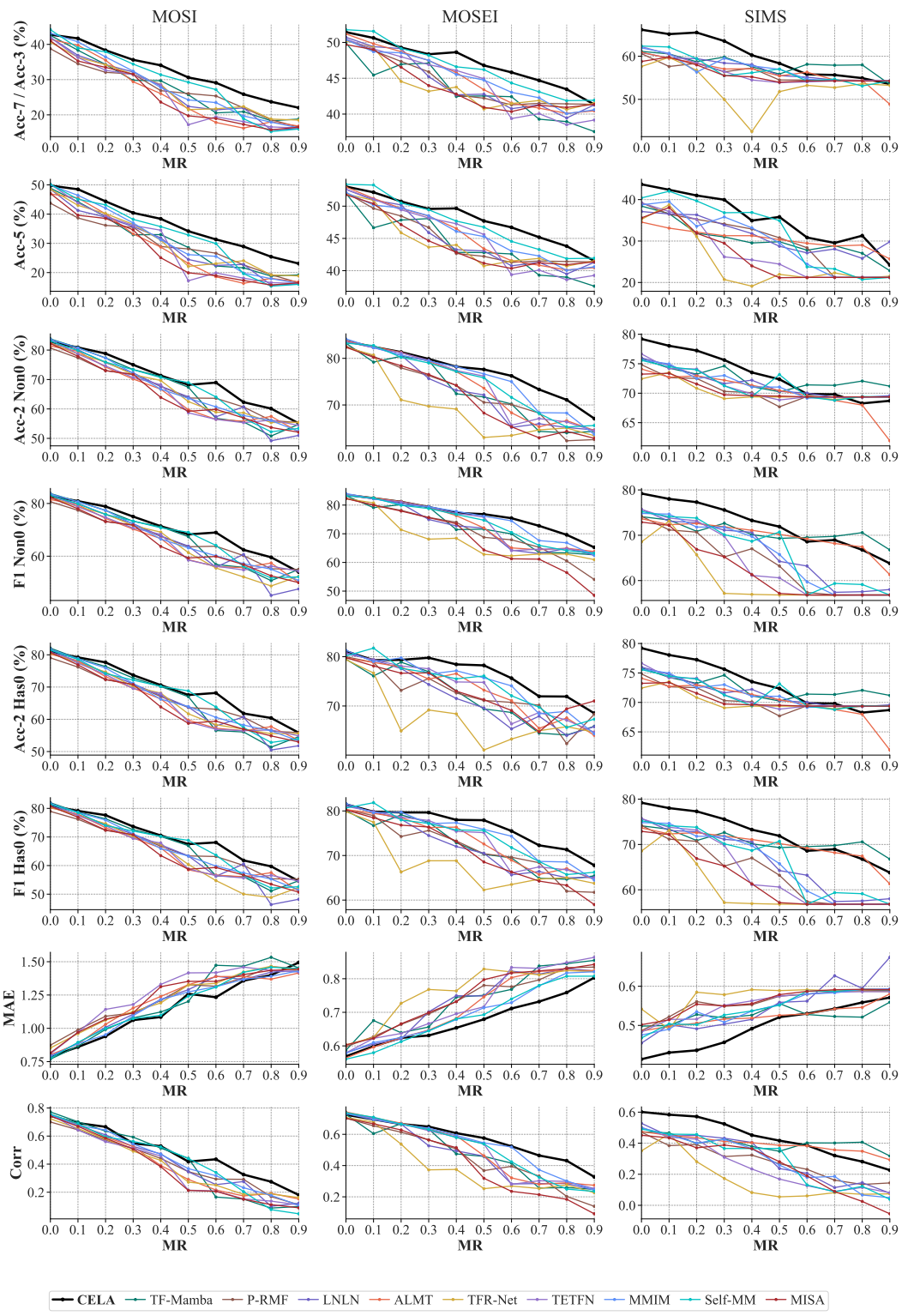


Figure 6: Performance across test missing rates for the constituent metrics on MOSI, MOSEI, and SIMS. Each curve reports the corresponding metric for one method across MR$= 0.0, \ldots, 0.9$. Lower MAE is better; higher values are better for the remaining metrics.

Matched-cell comparison. Table 14 compares CELA with the strongest rate-averaged baseline over the 30 matched $(\mathrm{MR}, \mathrm{seed})$ cells of each dataset.

Table 14: Matched-cell BGSS comparison with the strongest rate-averaged baseline.

Dataset	Reference baseline	ΔBGSS	Win / Tie / Loss
MOSI	Self-MM	+0.0255	23/0/7
MOSEI	MMIM	+0.0149	26/0/4
SIMS	TF-Mamba	+0.0206	21/0/9

CELA wins the majority of matched cells on all three datasets; the counts are descriptive rather than significance tests.

A.6 Detailed Ablation and Matched Baseline Analysis

The main paper reports MOSI ablation results averaged over the ten matched missing rates. Here we provide the exact configurations of the ablation variants, their rate-wise behavior, and an additional matched Baseline comparison that controls for the generic benefit of multi-view training exposure.

Table 15: Configurations of the MOSI ablation variants.

Variant	Configuration
w/o CESF	Replace CESF with concatenation-based fusion and disable the common-evidence branch and $\mathcal{L}_{\mathrm{CTA}}$.
w/o CMC	Set $\lambda_{\mathrm{cmc}} = 0$.
w/o FSA-align	Set $\lambda_{\mathrm{fa}} = 0$ while retaining full-view task supervision.
w/o FSA	Set $\lambda_{\mathrm{fs}} = 0$ and $\lambda_{\mathrm{fa}} = 0$.

All variants otherwise follow the same data, optimization, checkpoint selection, and evaluation protocol as CELA.

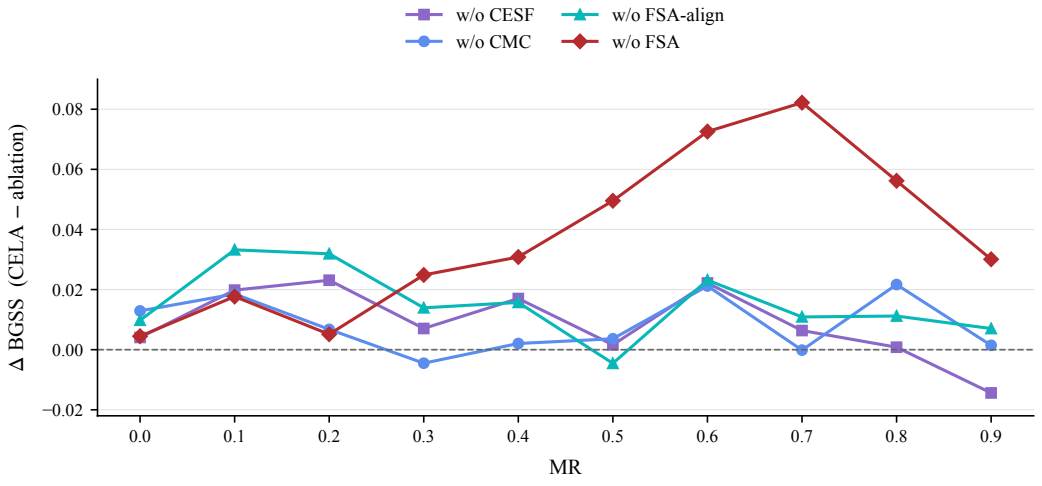


Figure 7: Rate-wise BGSS differences between CELA and its ablations on MOSI. Positive values favor CELA.

Removing both FSA terms produces the largest degradation over most moderate-to-high missing rates, whereas the effects of CESF and CMC vary more across missing rates. This rate dependence

is consistent with the different roles of the three components: CESF organizes modality-specific and common evidence within each observation, FSA coordinates incomplete representations with a task-supervised full-view reference, and CMC further regularizes variation across independently corrupted views.

Matched Baseline control. To separate the proposed design from the generic benefit of multi-view training, we compare CELA with the near-parameter-matched Baseline under FP32 software evaluation. Both methods use one clean supervised view and two independently corrupted views, with the same task supervision, optimization, checkpoint selection, and Test masks. The Baseline uses concatenation followed by a residual MLP and does not include FSA alignment, CMC, or CTA. Within the Fusion+Prediction scope, CELA and the Baseline contain 1,528,832 and 1,528,064 weights, respectively, differing by only 0.0502%.

Table 16: FP32 software-level matched comparison between CELA and the Baseline on MOSI. Values are mean±sample standard deviation over three seeds. Acc-2 and F1 use the Non0 convention; lower MAE is better.

MR	Method	BGSS	Acc-2	F1	MAE↓	Corr
0.1	CELA	0.6962 ± 0.0075	0.8059 ± 0.0093	0.8056 ± 0.0087	0.9232 ± 0.0272	$\mathbf{0.6935 \pm 0.0020}$
	Baseline	0.6951 ± 0.0095	$\mathbf{0.8089 \pm 0.0141}$	$\mathbf{0.8086 \pm 0.0133}$	$\mathbf{0.8647 \pm 0.0393}$	0.6890 ± 0.0386
0.5	CELA	$\mathbf{0.6002 \pm 0.0102}$	$\mathbf{0.7053 \pm 0.0189}$	$\mathbf{0.7045 \pm 0.0158}$	$\mathbf{1.1729 \pm 0.0479}$	$\mathbf{0.4592 \pm 0.0040}$
	Baseline	0.5764 ± 0.0200	0.6682 ± 0.0241	0.6687 ± 0.0252	1.2513 ± 0.0657	0.4232 ± 0.0467
0.9	CELA	$\mathbf{0.4904 \pm 0.0110}$	$\mathbf{0.5742 \pm 0.0213}$	$\mathbf{0.5702 \pm 0.0275}$	$\mathbf{1.3994 \pm 0.0271}$	$\mathbf{0.1931 \pm 0.0191}$
	Baseline	0.4653 ± 0.0168	0.5401 ± 0.0139	0.5258 ± 0.0179	1.4750 ± 0.0660	0.1586 ± 0.0215

At MR$= 0.1$, the two methods achieve nearly identical mean BGSS. The gap increases to $+0.0238$ at MR$= 0.5$ and $+0.0250$ at MR$= 0.9$, with corresponding improvements across the constituent metrics at these two rates.

Table 17: Seed-level BGSS for the FP32 matched Baseline comparison on MOSI. Δ denotes CELA minus Baseline.

MR	Seed	CELA	Baseline	ΔBGSS
0.1	1111	0.6875	0.6904	-0.0029
0.1	2222	0.7007	0.6889	$+0.0118$
0.1	3333	0.7004	0.7060	-0.0056
0.5	1111	0.5938	0.5757	$+0.0181$
0.5	2222	0.5948	0.5568	$+0.0381$
0.5	3333	0.6120	0.5968	$+0.0152$
0.9	1111	0.4869	0.4784	$+0.0085$
0.9	2222	0.5027	0.4464	$+0.0563$
0.9	3333	0.4815	0.4712	$+0.0103$

CELA wins all six seed-level comparisons at MR$= 0.5$ and MR$= 0.9$. Thus, under matched multi-view exposure and task supervision, its advantage becomes most evident under moderate-to-severe missingness.

A.7 Additional Analysis of the Number of Corrupted Views

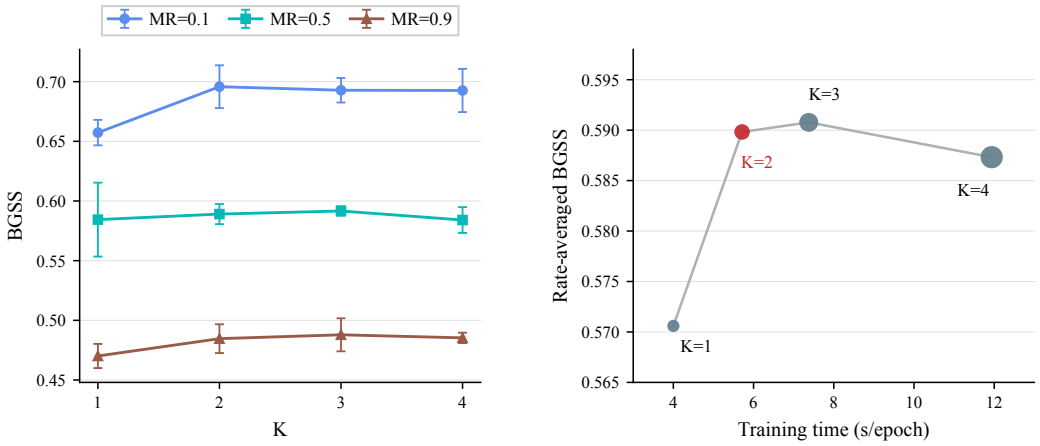
We evaluate $K \in \{1, 2, 3, 4\}$ corrupted training views on MOSI at MR$\in \{0.1, 0.5, 0.9\}$ with seeds 1111, 2222, and 3333. Within each fixed $(\mathrm{MR}, \mathrm{seed})$ pair, all K settings use the same Test-mask realization and validation-BGSS checkpoint selection.

Table 18: Seed-level BGSS for different numbers of corrupted views on MOSI.

K	MR	Seed 1111	Seed 2222	Seed 3333	Mean $\pm$ std
1	0.1	0.6524	0.6696	0.6499	0.6573 ± 0.0107
	0.5	0.5489	0.5983	0.6059	0.5844 ± 0.0310
	0.9	0.4610	0.4810	0.4684	0.4701 ± 0.0101
2	0.1	0.7044	0.7077	0.6752	0.6958 ± 0.0179
	0.5	0.5988	0.5844	0.5840	0.5890 ± 0.0084
	0.9	0.4709	0.4932	0.4899	0.4846 ± 0.0121
3	0.1	0.6812	0.7008	0.6964	0.6928 ± 0.0103
	0.5	0.5873	0.5926	0.5950	0.5916 ± 0.0039
	0.9	0.4798	0.4800	0.5039	0.4879 ± 0.0138
4	0.1	0.6851	0.6795	0.7132	0.6926 ± 0.0181
	0.5	0.5853	0.5726	0.5942	0.5841 ± 0.0108
	0.9	0.4885	0.4871	0.4803	0.4853 ± 0.0043

Seed-level performance. The rate-averaged BGSS values are 0.5706, 0.5898, 0.5908, and 0.5873 for $K = 1, 2, 3, 4$, respectively. The largest gain occurs from $K = 1$ to $K = 2$, while the difference between $K = 2$ and $K = 3$ is only 0.0010.

Changing K affects both training cost and the multi-view objectives: the incomplete-view task and FSA terms are averaged over the K corrupted views, while CMC operates over corrupted-view pairs and is absent for $K = 1$. Inference always uses a single incomplete view.



(a) Performance by MR (b) Performance and cost

Figure 8: Performance–cost trade-off for the number of corrupted views on MOSI. Left: BGSS across missing rates. Right: rate-averaged BGSS versus epoch time, with marker area proportional to peak memory.

Performance–cost trade-off. Although $K = 3$ gives the highest average BGSS, its gain over $K = 2$ is only 0.0010, while epoch time increases from 5.71 s to 7.38 s and peak memory increases by 1,099 MiB. We therefore use $K = 2$ as the default performance–cost trade-off.

A.8 Robustness Across Missingness Patterns

The main evaluation adopts independently sampled Bernoulli erasure at eligible Language tokens and Audio/Visual frames. We further examine whether the comparative behavior of the methods

changes when missingness is redistributed across modalities or organized into different temporal structures.

This analysis is evaluation-only on the 686 MOSI test utterances. For each method and seed, we freeze the checkpoint trained at MR$= 0.5$ and selected by validation BGSS; no model is retrained or reselected for the conditions below. We compare CELA with ALMT, LNLN, P-RMF, and TF-Mamba using seeds 1111, 2222, and 3333. Removed Language tokens are replaced by the BERT `[UNK]` token, whereas removed Visual and Audio frames are zeroed. The same semantic masks are paired across methods.

We consider three complementary families of stress tests. Whole-modality removal follows established robustness evaluations under partial and complete modality absence (Ma et al., 2022). Asymmetric modality missingness is motivated by recent studies of imbalanced missing rates across modalities (Nguyen et al., 2026). Structured temporal missingness follows prior evaluations of block-wise missing observations in sequential data (Peng et al., 2025; Tran et al., 2017). Based on these settings, we construct controlled variants that isolate modality allocation and temporal topology while keeping the overall missingness budget comparable. Their exact definitions are summarized in Table 19.

Table 19: MOSI stress-test conditions under different missingness patterns.

Family	Condition	Construction
Whole-modality removal	Full	No input position is removed.
	w/o Language	Remove all eligible Language tokens; special tokens and padding remain.
	w/o Visual	Remove all valid Visual frames.
	w/o Audio	Remove all valid Audio frames.
Asymmetric modality	Balanced	Bernoulli rates $(r_L, r_V, r_A) = (0.50, 0.50, 0.50)$.
	L-heavy	Bernoulli rates $(0.80, 0.35, 0.35)$.
	V-heavy	Bernoulli rates $(0.35, 0.80, 0.35)$.
	A-heavy	Bernoulli rates $(0.35, 0.35, 0.80)$.
Temporal structure	IID-Exact	Uniform sampling without replacement under a fixed removal budget.
	Independent block	One independently positioned contiguous block per modality.
	Independent 4-burst	Four temporal partitions with one local contiguous burst in each.
	Synchronized block	Modality blocks share the same normalized temporal position.
	Staggered block	Modality blocks occupy randomized early, middle, and late positions.

The asymmetric conditions preserve an average configured MR of 0.5 while redistributing missingness across modalities. For the temporal conditions, each sample and modality uses the same exact removal budget—half of its eligible positions, with unbiased rounding for odd sequence lengths. Therefore, the five temporal settings vary the arrangement of missing positions while holding their amount fixed. IID-Exact serves as the fixed-budget counterpart of the Bernoulli protocol used in the main experiments.

Figure 9 shows that CELA achieves the highest mean BGSS in all asymmetric and temporal conditions, as well as under Full, w/o Visual, and w/o Audio in the whole-modality experiment. Complete Language removal is the main exception: all methods degrade substantially, and CELA is not the strongest method in this condition.

For completeness, Table 20 reports the corresponding CELA metrics for all 13 conditions.

Sensitivity to modality allocation. The whole-modality and asymmetric tests reveal a clear dependence on which modality loses information. Removing Language decreases CELA from 0.6880 to 0.3984 BGSS, whereas removing Visual or Audio has a much smaller effect. A similar pattern

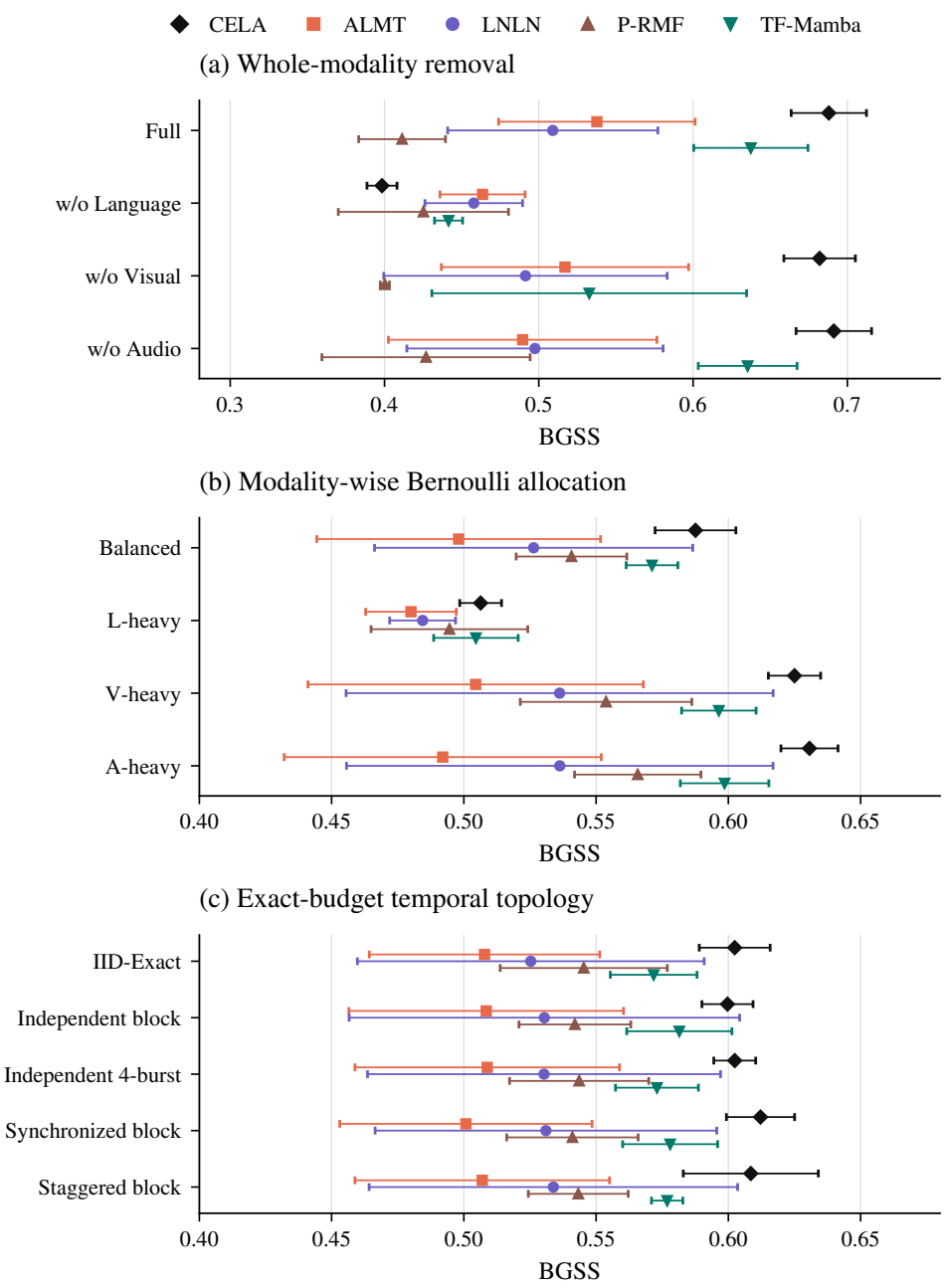


Figure 9: BGSS under different missingness patterns on MOSI. Results compare CELA, ALMT, LNLN, P-RMF, and TF-Mamba using frozen train-MR$=0.5$ checkpoints; error bars denote sample standard deviations across three seeds.

appears when the average configured missing rate is held at 0.5: relative to Balanced missingness, L-heavy missingness reduces BGSS by 0.0812, whereas V-heavy and A-heavy conditions increase it by 0.0374 and 0.0431, respectively.

Importantly, CELA remains the strongest of the five compared methods in all four asymmetric conditions. These results indicate that average MR alone does not fully characterize missingness difficulty: redistributing the same average corruption budget across modalities can substantially change performance.

Table 20: CELA performance under the MOSI missingness-structure stress tests. Values are mean ± sample standard deviation over three seeds. Indep. denotes Independent and Sync. denotes Synchronized. Acc-2 and F1 use Non0; lower MAE is better.

Condition	BGSS	Acc-2	F1	MAE↓	Corr
Whole-modality removal					
Full	0.6880 ± 0.0244	0.7937 ± 0.0288	0.7931 ± 0.0269	0.8926 ± 0.0470	0.6837 ± 0.0228
w/o Language	0.3984 ± 0.0097	0.4370 ± 0.0115	0.3122 ± 0.0183	1.6561 ± 0.0854	0.1335 ± 0.0489
w/o Visual	0.6820 ± 0.0232	0.7896 ± 0.0280	0.7892 ± 0.0262	0.9097 ± 0.0445	0.6814 ± 0.0169
w/o Audio	0.6912 ± 0.0245	0.7973 ± 0.0318	0.7961 ± 0.0302	0.8791 ± 0.0377	0.6837 ± 0.0228
Asymmetric modality missingness					
Balanced	0.5876 ± 0.0153	0.6855 ± 0.0170	0.6830 ± 0.0136	1.1962 ± 0.0334	0.4268 ± 0.0590
L-heavy	0.5063 ± 0.0079	0.5945 ± 0.0185	0.5896 ± 0.0229	1.4550 ± 0.0331	0.2337 ± 0.0567
V-heavy	0.6250 ± 0.0099	0.7277 ± 0.0130	0.7264 ± 0.0088	1.0834 ± 0.0318	0.5293 ± 0.0260
A-heavy	0.6307 ± 0.0108	0.7332 ± 0.0130	0.7316 ± 0.0091	1.0662 ± 0.0263	0.5301 ± 0.0290
Temporal missingness structure					
IID-Exact	0.6024 ± 0.0134	0.6951 ± 0.0250	0.6940 ± 0.0210	1.1542 ± 0.0353	0.4717 ± 0.0312
Indep. block	0.5997 ± 0.0097	0.6961 ± 0.0158	0.6953 ± 0.0140	1.1481 ± 0.0061	0.4869 ± 0.0329
Indep. 4-burst	0.6024 ± 0.0079	0.6992 ± 0.0248	0.6981 ± 0.0202	1.1416 ± 0.0236	0.4749 ± 0.0129
Sync. block	0.6122 ± 0.0129	0.7114 ± 0.0233	0.7104 ± 0.0208	1.1263 ± 0.0351	0.4968 ± 0.0123
Staggered block	0.6085 ± 0.0255	0.7144 ± 0.0289	0.7142 ± 0.0282	1.1382 ± 0.0493	0.4878 ± 0.0456

Sensitivity to temporally structured missingness. The temporal experiment isolates a different factor by keeping the number of removed positions fixed while changing their arrangement. CELA remains the strongest method in all five temporal conditions. Relative to IID-Exact, its paired BGSS changes are approximately -0.0027 for an independent block, $+0.0000$ for four bursts, $+0.0098$ for a synchronized block, and $+0.0061$ for staggered blocks.

Thus, within the tested fixed-budget settings, concentrating missing positions into blocks or bursts does not produce a systematic degradation. The observed variation across temporal topologies is also considerably smaller than the variation caused by shifting the corruption budget toward Language. Overall, these stress tests suggest that performance is more sensitive to *which modality* is missing than to the precise temporal organization of a fixed missingness budget.

Boundary under complete modality absence. CELA remains the strongest compared method across all asymmetric missingness conditions, including severe Language corruption. However, complete Language removal is substantially more challenging: all methods degrade sharply and CELA is no longer the strongest. This suggests that CELA is more effective at coordinating partially retained evidence than at fully compensating for the complete loss of a highly influential modality. Improving cross-modal compensation under complete modality absence remains an important direction for future work.

These experiments are MOSI-only and use frozen train-MR=0.5 checkpoints; they complement rather than replace the matched-rate evaluation in the main paper.

A.9 CIM Simulation and Physical Hardware Details

This section provides additional implementation details for the CIM-oriented evaluation in Section 4.7. We first describe the physical ReRAM platform and comparison setup, followed by the short- and long-term CIM simulations and the physical hardware evaluation.

HE-100 platform and comparison setup. The physical evaluation is conducted on the HE-100 ReRAM-based compute-in-memory (CIM) platform from the Helium system family. HE-100 integrates programmable control hardware, ReRAM array modules, and software interfaces for neural-network deployment. Each array adopts a 2T2R organization and provides a 576×128 logical weight space. The platform supports weight programming, readback, and matrix–vector computa-

tion through its software interface, enabling pretrained linear operators to be quantized and mapped onto the ReRAM arrays.

The CIM experiments compare CELA with a near-parameter-matched conventional Baseline. Both models use the same modality and BERT encoders and produce 256-dimensional modality representations. The Baseline replaces CELA's evidence-fusion pathway with a conventional concatenation-based fusion network while retaining the same $256 \to 256 \to 1$ prediction-head structure. The Baseline uses the same full- and incomplete-view task supervision as CELA, while FSA alignment, CMC, and CTA remain method-specific objectives of CELA. Thus, the comparison matches training exposure and task supervision while contrasting the proposed fusion and cross-view learning mechanisms.

Within the Fusion+Prediction mapping scope, CELA and the Baseline contain 1,528,832 and 1,528,064 mapped weights, respectively, differing by only 0.0502%.

Table 21: Mapped computation for CELA and the conventional Baseline.

Method	Mapped linear layers	Mapped weights	MACs / sample
CELA	22	1,528,832	5,464,576
Baseline	11	1,528,064	5,325,056

Both models are trained on MOSI at MR$= 0.5$ for 100 epochs. Each training sample contains one complete view and two independently corrupted views, and checkpoint selection uses validation BGSS only. The three CIM evaluations differ in mapping scope and nonideality model, as summarized in Table 22.

Table 22: CIM simulation and physical evaluation settings.

Study	Mapped scope	Numerical setting	Nonideality / repetitions
Short-term simulation	Fusion + Prediction	FP32 / W4A8	Multiplicative Gaussian weight perturbation with $\sigma \in \{0.1, 0.2, 0.3\}$; 50 paired trials.
Long-term simulation	Prediction Head	W4A8	Programming variation, conductance drift, and read noise; 50 seeds per condition and time point.
HE-100 hardware	Prediction Head	W4A8; three INT4 tiles	Three independent programming cycles; Complete and Block inputs reuse the same programmed weights within each cycle.

CIM simulation. In the short-term experiment, the Fusion+Prediction portion of each model is mapped using signed symmetric per-tensor W4 quantization,

$$s_W = \frac{\max |W|}{7}, \tag{20}$$

followed by rounding and clipping to $[-7, 7]$. Activations use symmetric A8 quantization with zero point zero and clipping to $[-127, 127]$. Activation scales are calibrated from the 99.9th percentile of valid validation activations, while biases, accumulation, and the remaining non-mapped arithmetic remain in FP32 (Chen et al., 2026).

Short-term weight variation is modeled multiplicatively as

$$W_{\text{noise}} = Q_{\text{W4}}(W) \odot (1 + \sigma\epsilon), \qquad \epsilon \sim \mathcal{N}(0, 1), \tag{21}$$

where the perturbation is sampled once for each mapped weight tensor and trial and remains fixed across all 686 MOSI test utterances. We evaluate $\sigma \in \{0.1, 0.2, 0.3\}$ using 50 paired trials for each stochastic condition.

Random and Block input conditions use matched realized removal budgets. Random missingness distributes removed positions over eligible inputs, whereas Block missingness concentrates them

into contiguous regions. As shown in Figure 4(a), CELA maintains a positive mean BGSS margin over the Baseline across the tested quantization and weight-noise conditions under both missingness patterns.

The long-term experiment maps only the prediction head $256 \to 256 \to 1$, containing 65,792 logical weights for either method. It uses the same W4A8 numerical precision and models programming variation, conductance drift, and read noise following the device nonideality model in Chen et al. (2026).

For target conductance G_T, programming variation is modeled by

$$\sigma_P = \max\left(-1.1731 G_T^2 + 1.9650 G_T + 0.2635,\ 0\right). \tag{22}$$

Given programmed conductance G_P, temporal drift follows

$$G_D(t) = G_P \left(\frac{t}{0.1}\right)^{-\nu}, \qquad \nu \sim \mathcal{N}(0.081, 0.015). \tag{23}$$

The read-noise scale is modeled as

$$\sigma_{\text{read}} = G_D Q \sqrt{\log\left(\frac{t + 250\,\text{ns}}{250\,\text{ns}}\right)}, \qquad Q = \min\left(\frac{0.0088}{G_T^{0.65}},\ 0.2\right). \tag{24}$$

Programming variation, drift, and read noise are evaluated individually and jointly using 50 stochastic seeds at 1 s, 3 days, 1 month, and 1 year. Figure 4(b) reports the combined temporal trajectory. At the one-year endpoint, CELA retains a positive BGSS margin over the Baseline under both Random (0.5436 vs. 0.5387) and Block (0.5343 vs. 0.5286) missingness, although the margin narrows under the accumulated long-term nonidealities.

Physical evaluation on HE-100. For the physical experiment, the prediction heads of CELA and the Baseline are mapped onto HE-100, while the modality encoders, multimodal fusion modules, and other upstream computations remain in software. The mapped prediction head has the form

$$256 \to 256 \to 1$$

and contains 65,792 logical weight coefficients excluding biases.

With the 576×128 logical array organization of HE-100, the first 256×256 linear layer is partitioned along its output dimension into two 256×128 tiles, while the final 256×1 layer uses a third tile:

$$256 \times 128, \qquad 256 \times 128, \qquad 256 \times 1.$$

Weights are quantized to W4 values in $[-7, 7]$, and activations use A8 values in $[-127, 127]$. Tile accumulation is performed using integer arithmetic, whereas digital bias addition, GELU, and tile-output concatenation remain outside the ReRAM computation. Activation scales are obtained from validation data only.

To characterize the efficiency of the physical implementation, we further compare GPU FP32 and HE-100 W4A8 execution under the same mapped prediction-head scope. Table 23 reports latency, effective compute throughput, power, and per-sample energy.

Table 23: Component-level efficiency comparison between GPU FP32 and physical HE-100 W4A8 execution under the same prediction-head scope.

Metric	GPU (L40, FP32)	CIM (HE-100, W4A8)
Latency (ms/sample)	0.02374	7.367
Effective throughput	5.543 GFLOP/s	0.01786 GOP/s
Average power (W)	95.392	0.135
Energy (mJ/sample)	2.265	0.995
Clock frequency (MHz)	2490	30

Under the same mapped computation scope, HE-100 substantially reduces power consumption and per-sample energy relative to GPU FP32 execution. Its higher latency is largely attributable to the much lower operating frequency of the current HE-100 platform (30 MHz versus 2490 MHz on the L40 GPU), rather than to an inherent limitation of the CIM computation paradigm. The present hardware therefore demonstrates a clear energy-efficiency advantage, while leaving substantial room for latency improvement through higher-frequency hardware operation.

All 686 MOSI test utterances are evaluated under Complete observations (MR$= 0$) and Block missingness with realized mean MR$= 0.498508$. Three independent HE-100 programming cycles are performed. Within each cycle, Complete and Block inputs reuse the same programmed weights, so the reported variation reflects programming-cycle variability.

The physical results are reported in Table 5. Under Block missingness, CELA improves BGSS from 0.5122 to 0.5324, Acc-2 by 4.36 percentage points, and F1 by 7.96 percentage points over the Baseline, while also obtaining lower MAE and higher Corr. The Complete-to-Block BGSS decrease is 0.0459 for CELA and 0.0583 for the Baseline.

Scope and real-world deployment. Although the evaluation extends beyond standard random missingness through structured missingness patterns, CIM-oriented nonidealities, and physical ReRAM execution, it does not yet fully reproduce real-world deployment conditions. The evaluated missingness patterns remain synthetically controlled rather than naturally occurring sensing, transmission, or synchronization failures. Moreover, the HE-100 experiment is component-level: only the prediction head is mapped to ReRAM-based CIM, while upstream encoding and fusion remain in software. Future work should therefore investigate naturally occurring missingness and end-to-end hardware deployment.