# Enhance-then-Balance Modality Collaboration for Robust Multimodal Sentiment Analysis

Supplementary Material

## A. Experimental Configuration

### A.1. Baselines

In this study, to systematically evaluate model performance under full-modality, missing-modality, and cross-task scenarios, we adopt a comprehensive suite of state-of-the-art baselines covering multimodal fusion, semantic disentanglement, robustness modeling, and imbalance mitigation.

- MulT [43]: proposes a Multimodal Transformer that models unaligned multimodal sequences using directional cross-modal attention, enabling interactions across modalities and time steps while capturing long-range dependencies.
- SelfMM [60]: proposes a self-supervised framework that generates unimodal labels to enable joint training of multimodal and unimodal tasks, using a dynamic weight-adjustment strategy to balance subtasks and enhance the capture of modality-specific differences.
- ConKI [61]: proposes a framework that injects domain-specific and general knowledge via contrastive learning to enhance multimodal sentiment prediction.
- ConFEDE [57]: proposes a unified framework that enhances multimodal sentiment representations by jointly performing contrastive representation learning and feature decomposition, separating modality-invariant and modality-specific features across text, video, and audio.
- CLGSI [58]: proposes sentiment-intensity-guided contrastive learning with a fusion mechanism to jointly capture common and modality-specific features for multimodal sentiment prediction.
- DEVA [54]: proposes a progressive fusion framework that converts audio-visual inputs into textualized sentiment descriptions and fuses them via a text-guided module to capture subtle emotional variations.
- EUAR [12]: proposes a Mixture-of-Experts framework with uncertainty-aware routing to handle noisy data in MSA, dynamically directing samples to experts with lower uncertainty to extract clearer features.
- GLoMo [70]: proposes a Global-Local Fusion framework that integrates local representations via modality-specific experts and combines them with global features through a guided fusion module, enhancing multimodal sentiment, humor, and emotion analysis.
- Semi-IIN [27]: proposes a semi-supervised network that captures intra- and inter-modal interactions using masked attention and gating, dynamically selects important features, and leverages unlabeled data to improve MSA.
- MCTN [35]: learns joint representations via modality translation with cycle-consistency, enabling robust sentiment prediction from a single modality.
- TransM [51]: proposes a Transformer-based fusion method that translates between modalities to learn joint representations, improving MSA.
- SMIL [28]: proposes a Bayesian meta-learning framework for multimodal learning with missing modalities during training and testing, achieving robust performance even with severely incomplete data.
- GCNet [26]: proposes a graph-based framework for incomplete multimodal conversation understanding, using Speaker and Temporal GNNs to capture dependencies and jointly optimizing classification and reconstruction.
- UMDF [21]: proposes a self-distillation framework with multi-grained crossmodal interaction and dynamic feature integration to handle uncertain missing modalities, producing robust multimodal representations for sentiment analysis.
- DMD [24]: proposes Decoupled Multimodal Distillation that separates each modality into modality-irrelevant and modality-exclusive spaces and applies a dynamic graph distillation unit to flexibly transfer crossmodal knowledge, enhancing emotion recognition.
- CorrKD [22]: proposes a knowledge distillation framework for uncertain missing modalities, using contrastive, prototype-guided, and response-disentangled strategies to reconstruct missing semantics and improve MSA.
- LNLN [66]: proposes language-dominated framework with dominant-modality correction and multimodal learning modules to enhance robustness and performance in MSA under noisy or missing data.

### A.2. Implementation Details

To ensure fair and reproducible comparisons with prior work, we follow the standard feature extraction settings widely adopted in multimodal sentiment/emotion analysis.

**Textual Modality.** We extract textual features using the $\texttt{BERT}_{\text{base}}$ model [7]. Each utterance is encoded through the 12-layer Transformer encoder, from which we take the contextualized hidden states of size 768 as the word-level representations. Following established IEMOCAP ERC protocols, we additionally incorporate 300-dimensional GloVe embeddings [34].

**Visual Modality.** We adopt the Facet toolkit to derive a 35-dimensional vector of facial action unit (AU) features for each frame. The AU descriptors capture fine-grained

facial muscle activations related to emotional expressions. For utterance-level features, the AU sequences are first temporally aligned with the text and audio streams and then aggregated using mean pooling.

**Acoustic Modality.** We use the COVAREP toolkit [6] to extract 74-dimensional low-level descriptors, including glottal source parameters, prosodic cues, spectral features, and other physiologically interpretable features. Frame-level features are resampled to match the temporal resolution of other modalities before being aggregated into utterance-level representations.

**Training Setup.** All hyperparameters in EBMC are tuned on the validation set. Unless otherwise specified, we set $\lambda_1, \lambda_2, \beta, \gamma, \eta = 0.1$ and $\zeta = 0.5$, which provides a stable balance among reconstruction, semantic enhancement, and modality collaboration objectives.

Each stage of the EBMC framework is trained independently, and we tailor the batch size and training schedule to the characteristics of each dataset. Unless otherwise specified, the learning rate is fixed at 0.0001, which we found to provide stable convergence across all datasets. For CMU-MOSEI, due to its large scale and substantial modality diversity, we adopt a batch size of 32. The full model is trained for 200 epochs, with 100 epochs allocated to each stage of EBMC. For CMU-MOSI, which is smaller but still rich in modality-specific variations, we use a batch size of 64 to accelerate training. The model is trained for 300 epochs, with 150 epochs per stage, ensuring adequate optimization in both the enhancement and collaboration phases. For IEMOCAP, given its limited size and higher speaker variability, we reduce the batch size to 16 to maintain stable gradient estimates. Each stage of the framework is trained for 150 epochs, providing a balanced schedule that avoids overfitting while ensuring sufficient convergence. All experiments are conducted on a workstation equipped with an NVIDIA RTX 4090 GPU (48GB). Training is implemented in PyTorch with mixed-precision acceleration enabled by default.

## A.3. Computational Cost and Complexity Analysis

The computational complexity of EBMC can be derived directly from the operations defined in MSD, CCE, EMC, and IMTD. Let $T_m$ and $d_m$ denote the temporal length and feature dimension of modality $m$, and let $h_m$ be the hidden width of the lightweight MLPs.

In MSD, the shared–specific decomposition requires two forward passes through modality-dependent MLPs, resulting in a cost of $O(T_m d_m h_m)$. The invariant alignment term based on InfoNCE further introduces a complexity of $O(|M|\, T_m d_m)$ due to pairwise similarities and softmax normalization, while the decorrelation of modality-specific components incurs an additional $O(|M|^2 d_m)$ complexity through cosine-similarity computations. In CCE, the enhancement network $G_m$ operates on compact aggregated vectors and therefore contributes only $O(d_m h_m)$ per modality. EMC consists solely of vector norms, entropy terms, energy differences, and energy gradients such as $\|z_m\|_2^2$, $H(p_{T_m})$, and $(E(m_i) - E(m_j))^2$, each of which is computed in $O(d_m)$ time. IMTD computes variances, confidence scores, and KL divergence terms in $O(T_m d_m)$ time without introducing additional nonlinear transformations.

Summing over all modalities yields the total computational complexity of EBMC:

$$O\left(\sum_m \left[T_m d_m h_m + |M|\, T_m d_m + |M|^2 d_m\right]\right), \quad (19)$$

which is dominated by the linear-time operations in MSD. Since EMC and IMTD consist exclusively of element-wise algebraic operations, they contribute negligible overhead relative to the backbone encoders and do not introduce attention-like quadratic costs at any stage of the framework.

Table 6. Comparison of training efficiency and parameters across different baselines on CMU-MOSI dataset. Please see Appendix A.4 for details.

| Model | Training Time | Params |
|---|---|---|
| SelfMM | 3.73h | 121,835,723 |
| GLoMo | 2.29h | 129,818,887 |
| EUAR | 1.16h | 110,436,422 |
| $\text{EBMC}_{\text{online}}$ | 1.04h | 112,320,746 |
| $\text{EBMC}_{\text{offline}}$ | 5min | 6,384,400 |

## A.4. Empirical Training Efficiency and Parameter Comparison

In addition to the theoretical analysis in Appendix A.3, we further report the empirical training efficiency of EBMC and other baselines in Tab. 6. All results are measured on the CMU-MOSI dataset by training each model for 300 epochs, ensuring a fair comparison across methods.

When EBMC is trained in the standard online setting—where raw inputs are loaded on-the-fly and each batch requires passing all modalities through BERT-based encoders—the end-to-end training takes approximately 1.2 hours, and the full system consists of 112,320,746 parameters, most of which originate from the Transformer encoders. In contrast, if multimodal features are pre-extracted offline and cached on disk, the training pipeline no longer depends on the heavy BERT encoders. Under this offline-feature configuration, EBMC reduces to only the lightweight MSD, CCE, EMC, and IMTD modules, resulting in a compact model with merely 6,384,400 parameters, and the total training time for 300 epochs drops dramatically to around 5 minutes.

This comparison highlights that the majority of computational overhead in multimodal sentiment models stems from online text encoding, and that EBMC itself remains lightweight and efficient when decoupled from the feature extractor.

# B. Theoretical Foundations of Energy-guided Modality Coordination (EMC)

This appendix provides a rigorous theoretical formulation of the proposed Energy-guided Modality Coordination module. We first establish its interpretation as a structured Energy-based Model (EBM), then analyze its connection and distinction from traditional gradient-balancing methods, followed by a dynamical perspective on the induced energy dynamics, and finally present the full gradients of the EMC objective.

## B.1. From Modality Energy to Energy-based Model

For each modality $m$, the main paper defines a modality-specific energy potential:

$$E(m) = \alpha \|z_m\|_2^2 + \beta\, \ell_m + \gamma\, u_m, \tag{20}$$

where $z_m$ is the latent representation, $\ell_m$ is the modality-specific loss, and $u_m$ denotes predictive uncertainty:

$$u_m = \mathbb{E}_i\left[H\left(p^i_{T_m}(y)\right)\right], \qquad H(p) = -\sum_y p(y)\log p(y). \tag{21}$$

**Joint Multimodal Energy.** Let $M$ denote the set of modalities. We define the global multimodal energy as:

$$E_{\text{joint}} = \sum_{m\in M} E(m) + \eta \sum_{m<m'} \big(E(m) - E(m')\big)^2, \tag{22}$$

where the second term encourages *energy equilibrium* across modalities by penalizing pairwise energy gaps.

The joint distribution of all modality representations then forms a Gibbs distribution:

$$p_\theta(\{z_m\}) = \frac{\exp(-E_{\text{joint}}(\{z_m\}))}{Z(\theta)}, \tag{23}$$

which establishes EMC as a structured EBM over multimodal representations.

**Energy-descent Dynamics.** Classical EBM learning often involves a Langevin-style update:

$$z^{t+1} = z^t - \lambda\, \nabla_z E_{\text{joint}}(z^t) + \omega^t, \qquad \omega^t \sim \mathcal{N}(0, \lambda I). \tag{24}$$

EMC adopts the deterministic gradient-descent component at the modality level:

$$\Delta z_m = -\lambda \frac{\partial E(m)}{\partial z_m}, \tag{25}$$

which can be interpreted as a single-step Langevin gradient flow on the energy manifold, restricted to the modality-specific potential $E(m)$. This establishes a principled EBM interpretation of the EMC training dynamics.

## B.2. Relation to Gradient-balancing Approaches

**Equivalence in Form.** Traditional gradient-balancing methods modify per-modality gradients via

$$g_m^{\text{new}} = w_m g_m, \tag{26}$$

where $g_m$ is the original gradient and $w_m$ is an explicitly designed weight. In contrast, EMC introduces an energy-gap loss:

$$\mathcal{L}_{gap} = \sum_{m<m'} \big(E(m) - E(m')\big)^2. \tag{27}$$

Taking the derivative w.r.t. $z_m$ gives

$$\frac{\partial \mathcal{L}_{gap}}{\partial z_m} = 2 \sum_{m'\neq m} \big(E(m) - E(m')\big) \frac{\partial E(m)}{\partial z_m}. \tag{28}$$

Letting $\bar{E} = \frac{1}{|M|}\sum_{k\in M} E(k)$, we observe

$$\sum_{m'\neq m} \big(E(m) - E(m')\big) = |M|\big(E(m) - \bar{E}\big), \tag{29}$$

which implies an implicit, energy-aware weight of the form

$$w_m^{\text{EMC}} \propto E(m) - \bar{E}. \tag{30}$$

Hence EMC behaves as an *energy-aware gradient modulator* whose weights are not manually chosen but emerge from the energy-gap objective.

**Fundamental Difference.** Unlike explicit gradient-balancing heuristics, EMC introduces a *global energy potential* and optimizes the system via its gradient flow. Thus, modality correction emerges *naturally* from the energy equilibrium rather than manual gradient manipulation.

Formally, EMC optimizes

$$\min_{\{z_m\}} \ \mathcal{L}_{EMC} = \mathcal{L}_{gap} + \delta \sum_{m\in M} \|\nabla_{z_m} E(m)\|^2, \tag{31}$$

which is a well-defined and fully differentiable objective. This grants EMC stronger theoretical grounding than ad-hoc gradient-balancing rules.

## B.3. Dynamical Analysis of EMC

Let $g_m = \nabla_{z_m} E(m)$ and again denote $\bar{E} = \frac{1}{|M|}\sum_{k\in M} E(k)$.

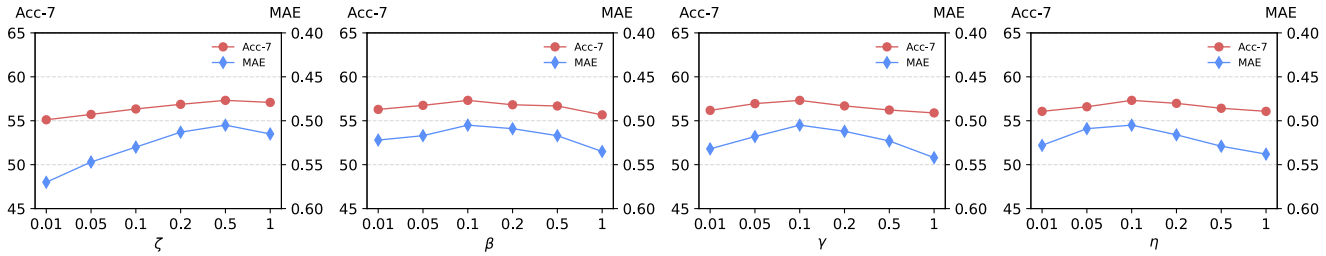


Figure 6. Hyperparameter sensitivity analysis of EBMC on the CMU-MOSEI dataset. The effects of varying $\beta$, $\gamma$, $\eta$, and $\zeta$ are evaluated according to the overall training objective defined in Eq. (17), while fixing the remaining hyperparameters to the values used in our main experiments. Please refer to Appendix C.1.

**Negative-feedback Structure.** From Eq. (28), the EMC-induced dynamics are proportional to

$$\Delta z_m \propto -\big(E(m) - \bar{E}\big)\, g_m, \tag{32}$$

up to a constant factor. This forms a classical negative-feedback system:

- if $E(m) > \bar{E}$ (weak or under-optimized modality), the factor $(E(m) - \bar{E})$ is positive and the gradient magnitude is effectively increased $\rightarrow$ *amplification*;
- if $E(m) < \bar{E}$ (dominant modality), the factor becomes negative and the effective gradient is reduced $\rightarrow$ *suppression*.

Thus EMC automatically stabilizes modality interactions by amplifying high-energy modalities and suppressing low-energy ones.

**Stability and Curvature Control.** The regularization term $\delta \sum_m \|\nabla_{z_m} E(m)\|^2$ introduces an additional dependence on the curvature of $E(m)$. Writing $H_m = \nabla^2_{z_m} E(m)$ for the Hessian, the local second-order behavior of $\mathcal{L}_{EMC}$ along $z_m$ contains terms of the form

$$2\|g_m\|^2 \quad \text{and} \quad 2\big(E(m) - \bar{E}\big)\, H_m, \tag{33}$$

up to higher-order derivatives. Intuitively, the first term penalizes large gradients and thus smooths sharp changes in the energy landscape, while the second term reduces curvature as modalities approach equilibrium, contributing to stable optimization dynamics.

**Convergence.** A stationary point of EMC satisfies

$$E(m_1) = \cdots = E(m_{|M|}), \qquad \nabla_{z_m} E(m) = 0, \quad \forall m. \tag{34}$$

At such points, all modalities lie in a balanced energy configuration and each representation reaches a local energy minimum, corresponding to a stable multimodal fusion state.

### B.4. Derivatives of EMC Components

**Energy Gradient.** From Eq. (20),

$$\nabla_{z_m} E(m) = 2\alpha z_m + \beta \nabla_{z_m} \ell_m + \gamma \nabla_{z_m} u_m. \tag{35}$$

**Derivative of Predictive Uncertainty.** For the entropy,

$$\frac{\partial H(p)}{\partial p(y)} = -\log p(y) - 1. \tag{36}$$

Thus:

$$\nabla_{z_m} u_m = \mathbb{E}_i \left[ \sum_y \big( -\log p^i_{T_m}(y) - 1 \big)\, \nabla_{z_m} p^i_{T_m}(y) \right]. \tag{37}$$

If $p^i_{T_m}(y)$ is produced by softmax logits $s^i_m(y)$, we have

$$\frac{\partial p^i_{T_m}(y)}{\partial s^i_m(y')} = p^i_{T_m}(y)\big(\mathbb{I}_{y=y'} - p^i_{T_m}(y')\big). \tag{38}$$

**Gradient of the Full EMC Objective.** The full EMC objective can be written as

$$\mathcal{L}_{EMC} = \sum_{m<m'} \big(E(m) - E(m')\big)^2 + \delta \sum_{m \in M} \|\nabla_{z_m} E(m)\|^2. \tag{39}$$

Differentiating w.r.t. $z_m$ yields

$$\nabla_{z_m} \mathcal{L}_{EMC} = 2 \sum_{m' \neq m} \big(E(m) - E(m')\big)\, \nabla_{z_m} E(m) + 2\delta\, H_m\, g_m, \tag{40}$$

where $g_m = \nabla_{z_m} E(m)$ and $H_m = \nabla^2_{z_m} E(m)$. The first term enforces pairwise energy equilibrium, while the second term regularizes the curvature of the energy landscape, jointly ensuring stable multimodal coordination.

## C. Additional Experiments and Analysis

### C.1. Sensitivity analysis

We evaluate the sensitivity of EBMC to the four coefficients in Eq. (17). Each coefficient is varied independently while keeping the others fixed, and we report Acc-7 and MAE on CMU-MOSEI. As shown in Fig. 6, the coefficients $\beta$ (CCE), $\gamma$ (EMC), and $\eta$ (IMTD) all exhibit a consistent trend: performance peaks at the moderate value of 0.1, achieving the best results. This suggests that a balanced

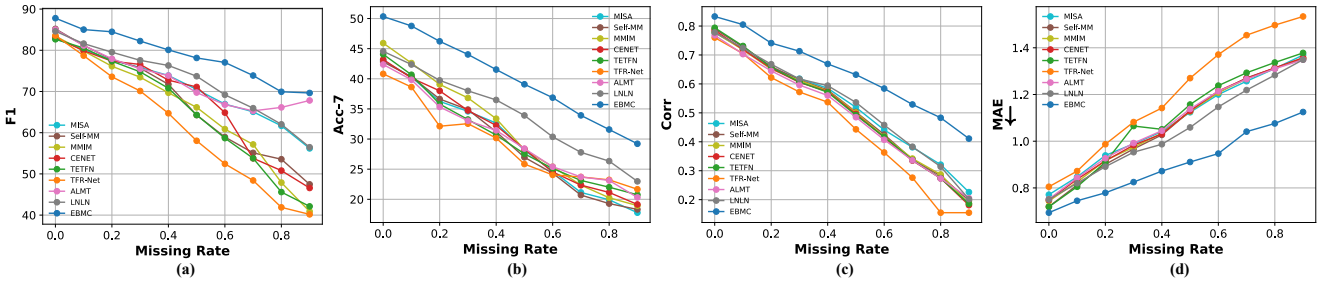


Figure 7. Performance curves on the CMU-MOSI dataset under increasing modality missing rates. The four subplots respectively report F1, Acc-7, Corr, and MAE across all compared models. Note that F1, Acc-7, and Corr are *higher-is-better* metrics, whereas MAE is a *lower-is-better* metric. Please refer to Sec. 4.3 (Q3) and Appendix C.3.

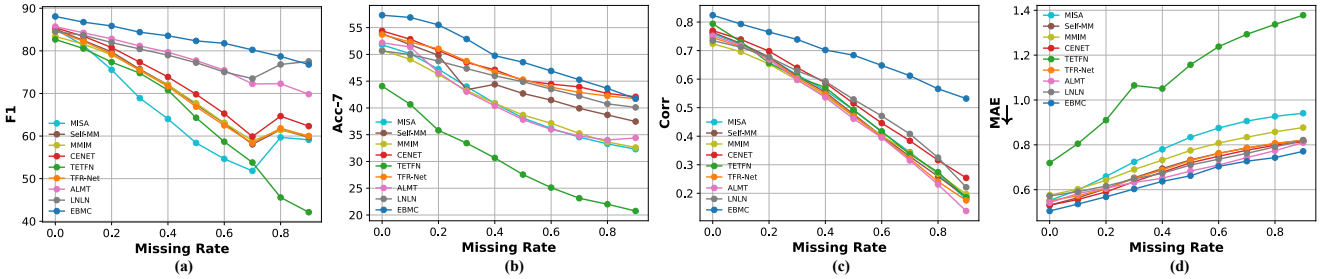


Figure 8. Performance curves on the CMU-MOSEI dataset under increasing modality missing rates. The four subplots respectively report F1, Acc-7, Corr, and MAE across all compared models. Note that F1, Acc-7, and Corr are *higher-is-better* metrics, whereas MAE is a *lower-is-better* metric. Please refer to Sec. 4.3 (Q3) and Appendix C.3.

level of regularization for these three objectives is most effective, whereas both under- and over-regularization lead to performance drops. The MSD coefficient $\zeta$ follows a different trend: performance improves monotonically up to $\zeta = 0.5$, where the same optimal score is achieved, and declines slightly at $\zeta = 1.0$. This indicates that MSD benefits from a stronger weight, but excessive emphasis becomes counterproductive. Overall, EBMC demonstrates stable behavior across a broad hyperparameter range, and the configuration used in the main experiments ($\beta = \gamma = \eta = 0.1,\ \zeta = 0.5$) closely matches the empirically optimal setting.

## C.2. Contribution of Energy-guided Modality Coordination

To better understand how each energy component contributes to EMC, we perform an ablation study by removing one term at a time (Table 7). Across all three metrics (F1, Acc-7, and MAE), every variant that removes a single component performs worse than the full EBMC model, showing that each term contributes a unique and irreplaceable effect.

Removing $|z_m|_2^2$ leads to a noticeable decline (e.g., Acc-7 drops from 57.32 to 56.62), indicating that the quadratic penalty helps stabilize the latent space and prevents uncontrolled shifts in modality embeddings. Eliminating the loss-coupling term $\ell_m$ results in the largest degradation (Acc-7=55.94), which aligns with its function of anchoring the energy landscape to task-relevant supervision. Omitting the uncertainty term $u_m$ also weakens performance, demonstrating that entropy-based calibration is valuable for distinguishing reliable from unreliable modalities during energy balancing. Overall, these results show that EMC is not a simple heuristic combination of losses; each component makes a substantive contribution to effective multimodal coordination.

Table 7. Ablation of EMC energy components on CMU-MOSEI. We report F1 (↑), Acc-7 (↑), and MAE (↓). Refer to Appendix C.2.

| | F1 (↑) | Acc-7 (↑) | MAE (↓) |
|---|---|---|---|
| EBMC | **86.23/88.07** | **57.32** | **0.505** |
| w/o $\|z_m\|_2^2$ | 85.54/87.52 | 56.62 | 0.518 |
| w/o $\ell_m$ | 85.05/86.89 | 55.94 | 0.541 |
| w/o $u_m$ | 85.36/87.16 | 56.40 | 0.527 |

## C.3. Extended Robustness Comparison under Random Missing Rates

Across Fig. 7 and Fig. 8, we observe a clear performance contrast between EBMC and existing baselines as the missing rate increases. Most baselines degrade rapidly once modality inputs become unreliable, revealing their strong dependence on fully observed data. In contrast, EBMC exhibits a noticeably smoother and more stable degradation trend. This resilience is attributed to two principles: the disentanglement of shared and modality-specific fac-

tors, which prevents excessive reliance on any single modality, and the enhancement–then–balance mechanism, which maintains cross-modal semantic cues even when certain modalities fail. As a result, EBMC avoids the majority-class collapse commonly seen in other methods under high missing rates and retains discriminative capability across the full corruption spectrum.

Another consistent observation is that nearly all models, including EBMC, perform better on CMU-MOSEI than on CMU-MOSI under the same missing rate. This difference arises from MOSEI's substantially larger scale and richer modality diversity, which provide stronger supervision and more stable temporal–semantic patterns. The additional data effectively buffers the impact of missing modalities, leading to improved robustness across models. EBMC benefits from these properties while still maintaining a clear advantage over competing approaches on both datasets, demonstrating that its robustness generalizes across data regimes rather than relying on dataset-specific artifacts.