E-MIND: CONDITIONAL ENERGY MATCHING FOR MULTIMODAL IMPUTATION AND DENOISING

Anonymous authors
Paper under double-blind review

ABSTRACT

Missing and corrupted modalities pose a common challenge for multimodal sentiment analysis: recovering task-relevant features from the remaining modalities. They nevertheless provide different starting points, as a missing modality offers no target observation, whereas a corrupted modality retains an imperfect one. We introduce E-MIND, a unified recovery framework that adapts Energy Matching to cross-modal conditional imputation and denoising. For each target modality, E-MIND jointly learns relative feature energies and recovery directions: contrastive energy learning encourages lower energies for real features than sampled negatives, while conditional flow regression supervises the negative energy gradient along noise-to-data interpolations. Conditioned on the remaining clean modalities, the resulting energy field generates missing features from Gaussian noise and refines corrupted features from their observations. Both modes reuse the same modality-specific energy parameters with different sampling settings and require no separate noise-specific training. Recovered and retained features are then fused for sentiment prediction. On CMU-MOSI, E-MIND achieves the highest weighted F1 among the compared methods at all seven nonzero nominal missing rates; on CMU-MOSEI, it achieves the highest seven-class accuracy in seven of eight settings. A recovery-loss ablation shows that combining the two objectives outperforms either alone, while separate single- and paired-modality corruption evaluations support reusing the learned fields for observation-based refinement. An extension of E-MIND to AVMNIST outperforms zero-fill and mean-fill baselines in classification accuracy, demonstrating the applicability of conditional energy recovery beyond sentiment analysis.

1 INTRODUCTION

Multimodal sentiment analysis combines language, acoustic cues, and facial behaviour to infer a speaker's affective state. Although these signals provide complementary information, their observations may be missing or corrupted. Missingness removes an input stream, whereas corruption leaves an available but unreliable observation. Both settings motivate recovering task-relevant features from the remaining modalities before fusion. This raises a common question: *How can a shared conditional recovery mechanism support sentiment prediction under missing or corrupted inputs?*

Existing approaches recover unavailable information through cross-modal translation, feature reconstruction, and conditional generation. Cyclic translation learns representations that can be used from a source modality alone (Pham et al., 2019), while modality imagination and graph completion reconstruct missing information from the available context (Zhao et al., 2021; Lian et al., 2023). IMDer further employs score-based diffusion for missing-modality recovery (Wang et al., 2023). Building on these approaches, we consider missing-feature generation and corrupted-feature refinement within a shared recovery framework. The two settings provide different starting points: a missing modality offers no target observation, whereas a corrupted modality provides an imperfect observation from which recovery can begin. In both cases, the remaining modalities supply context for recovering the target feature.

A conditional energy model offers a common mechanism for this recovery process. Its scalar value scores a candidate feature in its cross-modal context, while its negative gradient provides a direction for updating that feature. This formulation connects two aspects of recovery: *which candidates*

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

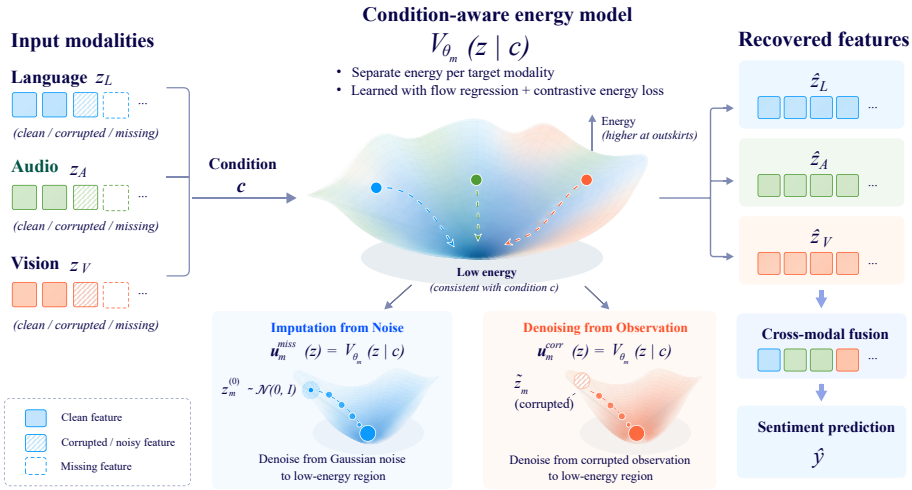


Figure 1: **Two recovery modes in E-MIND.** For each target modality, the same conditional energy field guides imputation from Gaussian noise and denoising from corrupted observations, using the remaining clean modalities as context. Recovered and retained features are fused for sentiment prediction. Missingness and corruption are evaluated separately.

receive lower energy and *how a candidate is updated*. We therefore seek to learn relative feature energies and recovery directions together, so that the same conditional potential can support both generation from noise and refinement from observations.

Based on this view, we introduce **E-MIND**, an energy-guided framework for multimodal imputation and denoising that adapts Energy Matching (Balcerak et al., 2026) to cross-modal conditional recovery. For each target modality, E-MIND learns a scalar energy conditioned on the available clean modalities. *Contrastive energy learning* encourages real target features to have lower energy than negative samples obtained from finite sampling chains under the same context (Hinton, 2002; Balcerak et al., 2026). In parallel, *conditional flow regression* directly supervises the negative energy gradient along noise-to-data interpolations. These objectives provide complementary supervision for a common potential: contrastive learning shapes relative energies, while flow regression trains the directions used for feature recovery.

The learned conditional energy field supports two recovery modes (Figure 1). For a missing modality, recovery starts from Gaussian noise and generates features conditioned on the observed modalities. For a corrupted modality, recovery starts from its observed feature and applies the same modality-specific field, conditioned on the remaining clean modalities. The two modes reuse the same energy parameters with mode-specific initializations, temperatures, and update counts, requiring no separate noise-specific training stage. Recovered features are then fused with retained observations for sentiment prediction.

Experiments on CMU-MOSI and CMU-MOSEI evaluate this shared recovery framework. Under random modality missingness, E-MIND achieves the highest weighted F_1 among the compared methods at all seven nonzero nominal missing rates on CMU-MOSI and the highest seven-class accuracy in seven of eight reported settings on CMU-MOSEI. The recovery-loss ablation supports combining the two learning signals: on CMU-MOSI at a nominal missing rate of 0.7, adding contrastive energy learning to flow regression raises F_1 from 60.7% to 71.4%, also exceeding the contrastive-only result of 70.3%. Separate evaluations across 13 single-modality and 55 paired-modality corruption conditions on CMU-MOSI examine the reuse of the learned fields for observation-based refinement. Beyond sentiment analysis, we extend E-MIND to AVMNIST, where it achieves higher classification accuracy than zero-fill and mean-fill baselines.

Our contributions are twofold:

- We introduce **E-MIND**, a cross-modal conditional recovery framework that adapts Energy Matching to jointly learn relative feature energies and recovery directions. The resulting

modality-specific energy fields are reused for both imputation and denoising without separate noise-specific training.

- We demonstrate the framework's effectiveness in separate missingness and corruption evaluations, with recovery-loss ablations supporting the combination of contrastive energy learning and flow regression. An extension to AVMNIST further demonstrates benefits beyond sentiment analysis.

2 Related Work

From fusion to incomplete-input prediction. Sentiment fusion captures cross-modal interactions (Zadeh et al., 2017; Liu et al., 2018; Zadeh et al., 2018a; Tsai et al., 2019) while preserving shared and modality-specific information (Hazarika et al., 2020; Yu et al., 2021; Han et al., 2021). Pretrained language integration and task sharing further exploit nonverbal cues and affective knowledge (Rahman et al., 2020; Hu et al., 2022). Missing inputs shift the problem from combining evidence to compensating for its absence (Wu et al., 2024). Modality dropout and meta-learning encourage prediction from incomplete observations (Neverova et al., 2015; Ma et al., 2021), while translation and reconstruction recover unavailable information (Pham et al., 2019; Zhao et al., 2021; Lian et al., 2023; Yuan et al., 2021). Recovery quality also depends on alignment and semantic consistency (Han et al., 2022; Zeng et al., 2022; Sun et al., 2023). Distillation and hierarchical learning strengthen incomplete representations (Li et al., 2024a;b); prompting and parameter-efficient adaptation encode missingness within existing networks (Guo et al., 2024; Reza et al., 2024). Language-guided correction and fusion exploit the dominant modality (Zhang et al., 2024; Li et al., 2025). Together, these approaches motivate recovery that preserves task-relevant evidence before fusion.

Generative feature recovery. Variational models infer missing modalities through products or mixtures of experts (Wu & Goodman, 2018; Shi et al., 2019; Sutter et al., 2021), while adversarial imputation distinguishes generated from observed components (Yoon et al., 2018). Diffusion models instead learn iterative denoising dynamics (Ho et al., 2020; Song & Ermon, 2019; Song et al., 2020), supporting conditional time-series imputation and image completion (Tashiro et al., 2021; Lugmayr et al., 2022). IMDer brings conditional score-based recovery to affective features (Wang et al., 2023), and GSDNet incorporates graph spectral structure (Shou et al., 2025). E-MIND retains conditional recovery but learns a scalar energy whose gradients support both noise-initialized imputation and observation-initialized refinement.

Energy and flow learning. Denoising connects reconstruction to density scores (Vincent, 2011; Alain & Bengio, 2014; Hyvärinen & Dayan, 2005). Contrastive objectives offer alternative routes to energy learning (Hinton, 2002; Gutmann & Hyvärinen, 2010; Du et al., 2020), enabling generation, restoration, and joint classification (Du & Mordatch, 2019; Grathwohl et al., 2019). Flow-based methods learn transport directions along interpolating paths (Lipman et al., 2022; Liu et al., 2022; Albergo & Vanden-Eijnden, 2022). Energy Matching combines flow regression with contrastive learning in a time-independent potential (Balcerak et al., 2026); we adapt this objective to cross-modal conditioning. Energy-Based Transformers train through prediction optimization (Gladstone et al., 2026); our sentiment loss instead acts on detached recovered features. Our straight-line target also differs from Equilibrium Matching's endpoint-stationary construction (Wang & Du, 2025). The resulting conditional field is reused across imputation and denoising without separate noise-specific training.

3 The Proposed Method

E-MIND learns a conditional energy model for each target modality and reuses its parameters for both imputation and denoising (Figure 2). The negative energy gradient guides feature recovery from Gaussian noise when a modality is missing, or from the observed feature when it is corrupted, with temperature settings tailored to each recovery mode. Recovered and retained features are then fused for sentiment prediction.

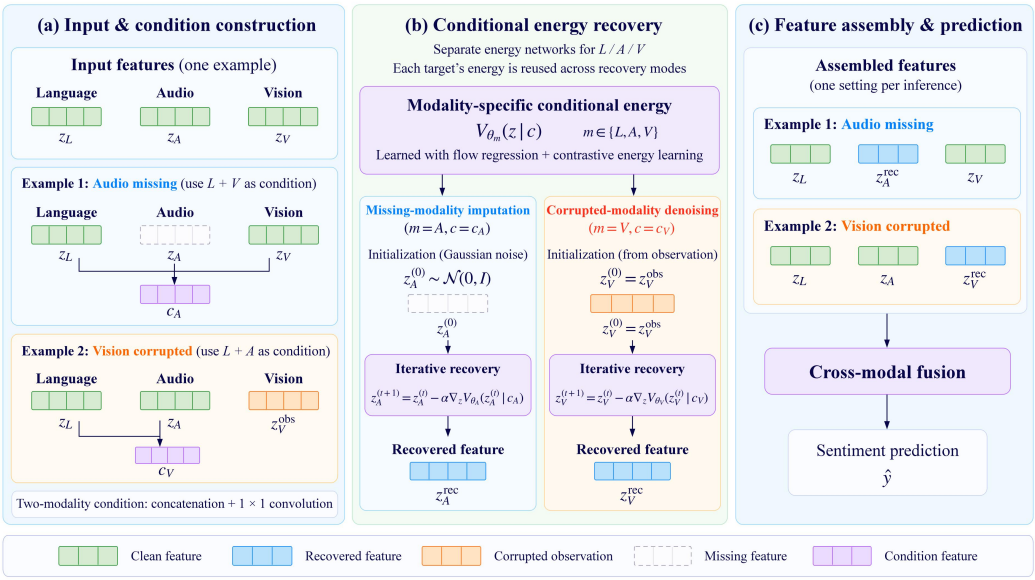


Figure 2: **Overview of E-MIND.** Learned conditional energy gradients guide iterative feature recovery: missing features are generated from Gaussian noise, while corrupted features are denoised from their observations. Recovered and retained features feed cross-modal fusion and sentiment prediction. Missingness and corruption are evaluated separately.

3.1 Problem Formulation

Given a multimodal utterance $x = (x_L, x_A, x_V)$, our task is to predict its sentiment score $y \in [-3, 3]$ from language, audio, and vision, indexed by $\mathcal{M} = \{L, A, V\}$. Encoders and temporal projections map the inputs to aligned feature sequences $z_m \in \mathbb{R}^{T \times d}$. When all modalities are observed and clean, their features can be directly fused for prediction. We consider two forms of degraded input: missing modalities, whose observations are unavailable, and noisy modalities, whose observations are present but corrupted. Our goal is to impute the missing features and denoise the noisy features before fusion.

For missing modalities, let $O \subseteq \mathcal{M}$ index the observed modalities and $H = \mathcal{M} \setminus O$ the missing ones. A plausible feature sampled independently of the utterance need not agree with its observed content. Imputation thus targets $p_{\text{data}}(z_m \mid z_O)$ for $m \in H$, where $z_O = \{z_j : j \in O\}$ provides the semantic context.

For noisy modalities, let $C \subseteq O$ index the modalities designated for denoising, and let $\widetilde{z}_m$ denote their observed features. The target is the clean feature z_m, conditioned on both its noisy observation and the remaining clean modalities, as expressed by $p_{\text{data}}(z_m \mid \widetilde{z}_m, z_{O \setminus C})$ for $m \in C$. We assess the recovered features through downstream sentiment prediction. We evaluate missingness and modality noise separately: $C = \varnothing$ and $\varnothing \neq O \subsetneq \mathcal{M}$ for missingness, and $O = \mathcal{M}$ for modality noise.

To address both forms of degradation, we learn conditional energy models that guide feature recovery using the available cross-modal context. The following sections define these models and explain how their gradients support imputation and denoising. In either case, the recovered features $\widehat{z}_m$ are combined with the retained clean features for prediction:

$$\bar{z}_m = \begin{cases} z_m, & m \in O \setminus C, \\ \widehat{z}_m, & m \in H \cup C, \end{cases} \qquad \widehat{y} = f_\phi(\bar{z}_L, \bar{z}_A, \bar{z}_V), \tag{1}$$

where f_ϕ performs cross-modal fusion and sentiment prediction. Fully observed training utterances supervise artificially hidden modalities; clean target features and sentiment labels are unavailable to recovery at inference.

3.2 CONDITIONAL ENERGY MODEL

The observed modalities provide the semantic context against which a candidate feature is evaluated. We denote this context by c_S, using $S = O$ for missingness and $S = O \setminus C$ for corruption. For imputation, a single observed modality supplies its projected feature sequence directly; when two modalities are observed, their features are concatenated along the channel dimension and mapped back to width d by a learned 1×1 convolution.

For each target modality m, a conditional Transformer maps the candidate sequence z and context c_S to a scalar energy $V_{\theta_m}(z \mid c_S)$. Candidate and context features are embedded separately with temporal positional encodings, and cross-attention allows the candidate representation to incorporate the observed context. Temporal average pooling and an energy head then produce one scalar per utterance. Each modality has its own energy parameters, which are reused for imputation and denoising of that modality. Targets are recovered independently from the observed context, without using generated features to condition other targets. The negative gradient with respect to the candidate sequence defines the recovery field:

$$v_{\theta_m}(z \mid c) = -\nabla_z V_{\theta_m}(z \mid c). \tag{2}$$

The network takes no interpolation or sampling time as input, so this field is autonomous. Holding the context fixed, we use it to update the candidate feature in either recovery mode.

3.3 MISSING-MODALITY GENERATION

For a missing modality, the energy field generates a feature sequence conditioned on the observed modalities. For each $m \in H$, initialize $z_m^{(0)} \sim \mathcal{N}(0, I)$ and hold c_O fixed. Euler–Maruyama updates give

$$z_m^{(k+1)} = z_m^{(k)} - h\nabla_z V_{\theta_m}(z_m^{(k)} \mid c_O) + \sqrt{2\epsilon(s_k)h}\,\eta_k, \qquad \eta_k \sim \mathcal{N}(0, I), \tag{3}$$

where $h > 0$ is the step size and $s_k = kh$. Following Energy Matching, the temperature schedule

$$\epsilon(s) = \begin{cases} 0, & s < \tau, \\ \epsilon_{\max}(s - \tau)/(1 - \tau), & \tau \le s < 1, \\ \epsilon_{\max}, & s \ge 1 \end{cases} \tag{4}$$

transitions from deterministic movement to stochastic exploration, with $0 \le \tau < 1$. The terminal iterate gives $\widehat{z}_m = z_m^{(K)}$ after K updates.

3.4 ENERGY-GUIDED DENOISING

The denoising procedure uses the corrupted feature as its initialization and applies the same learned energy field, conditioned on the remaining clean modalities. For each $m \in C$, initialize $z_m^{(0)} = \widetilde{z}_m$ and hold the clean context $c = c_{O \setminus C}$ fixed. The learned energy gradient then guides iterative denoising:

$$z_m^{(k+1)} = z_m^{(k)} - h\nabla_z V_{\theta_m}(z_m^{(k)} \mid c) + \sqrt{2\epsilon_d h}\,\eta_k, \qquad \eta_k \sim \mathcal{N}(0, I). \tag{5}$$

Here $h > 0$ is the step size and $\epsilon_d \ge 0$ is a constant temperature. At each step, the negative energy gradient provides a correction based on the current feature and its cross-modal context; $\epsilon_d = 0$ gives deterministic gradient descent, while $\epsilon_d > 0$ adds stochastic exploration. After K_d updates, we use $\widehat{z}_m = z_m^{(K_d)}$ for prediction. The energy is evaluated without a condition if no clean modality remains. The corrupted set is supplied to the procedure.

Thus, imputation and denoising reuse the energy parameters of each target modality while using different initializations, temperature settings, and numbers of updates. The noisy observation enters denoising through initialization, while the learned conditional field guides subsequent recovery; this reuse requires no separate noise-specific training stage. Appendix C analyzes autonomous flow regression and discusses the scope of its transport and equilibrium interpretation.

Algorithm 1 in Appendix A summarizes the two separately evaluated recovery modes. In each mode, the conditioning features remain fixed throughout recovery, and the recovered features are combined with the retained features for sentiment prediction.

3.5 Multimodal Fusion and Training Objectives

Recovered and retained features pass through pairwise cross-modal attention and temporal fusion (Tsai et al., 2019), followed by a residual prediction head. To learn the field used by both recovery modes, we combine flow regression, which supervises update directions along noise-to-data interpolations, with contrastive energy learning, which encourages lower energy for target features than for finite-chain negative samples. Sentiment supervision trains the predictor on the resulting feature representations.

Conditional flow regression. Following Energy Matching (Balcerak et al., 2026), we train the energy gradient using noise-to-data interpolation. For a hidden target z_m, draw independent $\xi \sim \mathcal{N}(0, I)$ and $t \sim \mathcal{U}[0, 1]$, and define

$$z_t = (1 - t)\xi + tz_m, \qquad u = z_m - \xi. \tag{6}$$

The flow loss is

$$\mathcal{L}^m_{\mathrm{FM}} = \mathbb{E}\left[w(y) \, \frac{\|v_{\theta_m}(z_t \mid c_O) - u\|_F^2}{Td} \right]. \tag{7}$$

Here $w(y) = 1$ gives uniform weighting; $w(y)$ can also encode optional sentiment-intensity weighting. Our implementation uses independent noise–data pairs without optimal-transport assignment and fits an autonomous approximation to the target velocities.

Contrastive energy learning. We also train the potential to distinguish target features from finite-chain negative samples (Hinton, 2002; Balcerak et al., 2026):

$$\mathcal{L}^m_{\mathrm{CD}} = \frac{1}{B} \sum_{i=1}^{B} \left[V_{\theta_m}(z_{m,i} \mid c_i) - V_{\theta_m}(\mathrm{sg}[z^-_{m,i}] \mid c_i) \right], \tag{8}$$

where B is the batch size, c_i is the observed context, and sg stops gradients through the negative chain. Half the chains, rounded down, start from target features and the remainder from Gaussian noise. The negative chains use the energy-driven updates introduced in Section 3.3.

Overall objective and optimization. Combining the sentiment loss with the two energy-learning losses gives

$$\mathcal{L} = |\hat{y} - y| + \lambda_F \sum_{m \in H} \mathcal{L}^m_{\mathrm{FM}} + \lambda_C \sum_{m \in H} \mathcal{L}^m_{\mathrm{CD}}. \tag{9}$$

Although these losses appear in a single objective, recovered features are detached before prediction, so sentiment supervision does not update the energy parameters θ_m through sampling. The flow loss, in contrast, differentiates through the energy gradient to train the recovery field. The weights λ_F and λ_C balance the recovery losses, with the contrastive term set to zero during a flow-only warm-up. Complete inputs bypass imputation and contribute no recovery losses.

Algorithm 2 in Appendix A combines these objectives in one training loop. Artificially hidden modalities provide targets for flow and contrastive learning, while their recovered features are passed to the sentiment predictor with the gradient separation described above. Training on artificially hidden modalities therefore supplies the conditional energy models reused by both recovery modes in Algorithm 1.

4 Experiments

4.1 Datasets and Implementation Details

Datasets. We conduct experiments on CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018b), two standard benchmarks of multimodal sentiment in online opinion videos. CMU-MOSI contains 2,199 utterance clips: 1,284 for training, 229 for validation, and 686 for testing. CMU-MOSEI contains 22,856 clips, split into 16,326 training, 1,871 validation, and 4,659 test samples. Each clip provides language, audio, and visual inputs. DeBERTa-v3-large (He et al., 2021) produces 1024-dimensional language features. For the visual modality, Facet extracts 35 facial action-unit features from each video frame. COVAREP extracts 74-dimensional acoustic features. Modality-specific projections map all three streams into a shared latent space.

Evaluation metrics. CMU-MOSI and CMU-MOSEI comprise YouTube videos in which speakers express their opinions. Both datasets provide human-annotated sentiment scores, whereas CMU-MOSEI additionally includes categorical emotion labels. We focus on the sentiment score, which ranges from -3 (highly negative) to 3 (highly positive) and includes seven ordered levels from highly negative to highly positive. We report seven-class accuracy (ACC_7), binary accuracy (ACC_2), and weighted F_1. For seven-class evaluation, predictions are clipped to $[-3, 3]$ and rounded to the nearest integer. For ACC_2 and weighted F_1, neutral labels are excluded and the remaining scores are binarized by sign. Main-text tables report metrics as percentages and metric differences in percentage points.

Implementation details. We evaluate robustness under two protocols: random modality missingness and fixed corruption combinations. Under *random missingness*, the available modality set varies across samples, allowing one or more modalities to be unavailable for each sample. We quantify the overall missingness using the empirical missing rate, defined for N samples and $M = |\mathcal{M}| = 3$ modalities as

$$\mathrm{MR}_{\mathrm{emp}} = 1 - \frac{\sum_{i=1}^{N} m_i}{NM}, \qquad 1 \le m_i \le M, \tag{10}$$

where m_i is the number of available modalities for sample i. We require at least one modality to remain available, giving $\mathrm{MR}_{\mathrm{emp}} \le (M-1)/M = 2/3$. For three modalities, we evaluate nominal MR values from 0.0 to 0.7 in increments of 0.1; the value 0.7 approximates this upper bound rather than representing a literal 70% removal of modality slots.

Under the fixed corruption-combination protocol, we derive corrupted conditions from the original CMU-MOSI test split rather than introducing a separate dataset. The evaluation includes 13 single-modality corruptions and 55 paired-modality combinations: 19 audio–vision, 20 text–audio, and 16 text–vision settings. The corrupted modalities are designated before refinement. Appendix D reports the condition-specific results.

4.2 Comparison with state-of-the-art methods

To benchmark E-MIND under random modality missingness, we compare it with methods that represent complementary strategies for incomplete multimodal learning. DCCA and DCCAE provide correlation-based baselines without explicitly modeling modality recovery, whereas MCTN, MMIN, GCNet, and IMDer reconstruct unavailable information through translation, imagination, graph completion, or conditional generation, respectively (Andrew et al., 2013; Pham et al., 2019; Zhao et al., 2021; Lian et al., 2023; Wang et al., 2023). We also include MOMKE and TMDC as additional comparison methods (Xu et al., 2024; Zhuang et al., 2026). Table 1 summarizes F_1 and ACC_7 across eight nominal MR settings on CMU-MOSI and CMU-MOSEI.

Quantitative results. Table 1 shows that E-MIND achieves the best F_1 across all eight missing-rate settings on CMU-MOSI. This advantage extends to substantial missingness: at MR 0.6, E-MIND exceeds the strongest baseline, IMDer, by 3.9 percentage points. These results highlight its ability to maintain sentiment polarity prediction with limited modality information.

On CMU-MOSEI, E-MIND's advantage is most evident in fine-grained sentiment prediction, with the highest ACC_7 in seven of eight settings. At the highest nominal MR of 0.7, it achieves 49.8%, outperforming the strongest baseline by 2.2 percentage points. This result complements the F_1 gains on MOSI, showing that E-MIND also supports sentiment-intensity prediction under substantial modality missingness.

4.3 Recovery under corrupted observations

We evaluate E-MIND under audio, visual, and textual corruptions, including environmental noise, silence, blur, black frames, pixel noise, and text perturbations. We apply refinement to the designated corrupted modalities and compare sentiment prediction with the baseline without denoising. Appendix D provides the condition-specific results.

Figure 3 summarizes the single-modality comparison using the metric most consistently improved by denoising. F_1 increases in all 13 conditions, with a mean gain of 1.64 percentage points. The

Table 1: Results under random modality missingness. Entries report F_1/ACC_7 (%). Bold and underline indicate the best and second-best values. MR: nominal missing rate.

Method	Missing rate								Reported average
	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	
CMU-MOSI									
DCCA	75.4/30.5	72.2/28.0	69.1/26.8	65.2/25.7	62.0/24.2	59.9/21.6	57.3/21.2	56.0/20.4	64.6/24.8
DCCAE	77.4/31.2	74.7/28.1	71.9/27.6	66.7/25.8	62.8/24.2	61.3/23.0	58.5/20.9	57.4/20.6	66.3/25.2
MCTN	81.5/43.4	78.5/39.8	75.7/38.5	71.2/35.5	67.6/32.9	64.8/31.2	62.5/29.7	59.0/27.5	70.1/34.8
MMIN	84.4/44.8	81.8/41.2	79.1/38.9	76.2/36.9	71.6/34.9	66.5/32.2	64.0/29.1	61.0/28.4	73.1/35.8
GCNet	85.1/44.9	82.3/42.1	79.5/40.0	77.2/38.2	74.4/36.6	69.8/33.9	66.7/29.8	65.4/28.1	75.1/36.7
IMDer	85.6/45.3	84.8/44.8	83.4/**44.3**	81.0/**42.5**	78.5/**39.7**	75.9/**37.9**	74.0/**35.8**	71.2/**33.4**	79.3/**40.5**
MOMKE	88.1/32.0	83.5/32.8	81.4/33.7	77.6/31.0	76.1/29.7	73.8/25.9	65.7/25.0	65.9/23.5	76.5/29.2
TMDC	85.4/35.7	83.2/35.2	79.4/11.6	77.6/31.0	68.5/11.6	72.9/27.6	67.4/24.4	67.0/23.8	75.2/25.1
E-MIND	**89.5/47.5**	**88.4/45.8**	**85.3**/44.0	**81.2**/41.4	**81.3**/37.2	**77.0**/34.3	**77.9**/34.3	**71.4**/31.8	**81.5**/39.5
CMU-MOSEI									
DCCA	80.9/47.7	77.3/46.2	74.0/45.1	71.2/43.6	69.4/43.1	65.4/42.5	63.1/42.4	61.0/42.1	70.3/44.1
DCCAE	81.2/48.2	78.3/46.9	75.4/46.3	72.2/45.6	70.0/44.0	66.4/43.3	63.2/42.9	62.6/42.5	71.2/45.0
MCTN	84.2/51.2	81.6/49.8	78.7/48.6	76.2/47.4	74.1/45.6	72.6/45.1	71.1/43.8	70.5/43.6	76.1/46.9
MMIN	84.2/52.4	81.3/50.6	78.8/49.6	75.5/48.1	72.6/47.5	70.7/46.7	70.3/45.6	69.5/44.8	75.4/48.2
GCNet	85.1/51.5	82.1/51.2	79.9/50.2	76.8/49.2	74.9/48.0	73.2/46.7	72.1/45.1	70.4/44.5	76.8/48.3
IMDer	85.1/53.4	84.6/53.1	82.4/52.0	80.7/51.3	78.1/50.0	77.4/**49.2**	75.5/48.5	74.6/47.6	79.8/50.6
MOMKE	87.1/44.0	84.7/46.6	83.6/45.5	81.8/43.8	79.4/40.8	77.7/40.0	75.9/36.9	73.5/36.1	80.5/41.7
TMDC	87.0/46.3	85.6/38.6	**84.9**/37.0	**83.0**/37.5	**82.6**/36.0	**81.7**/40.6	**80.7**/40.0	**80.4**/38.5	**83.2**/39.3
E-MIND	**87.6/55.0**	**86.4/55.6**	83.3/**52.4**	81.4/**52.1**	80.0/**51.2**	76.3/48.9	75.0/**48.9**	75.6/**49.8**	80.7/**51.7**

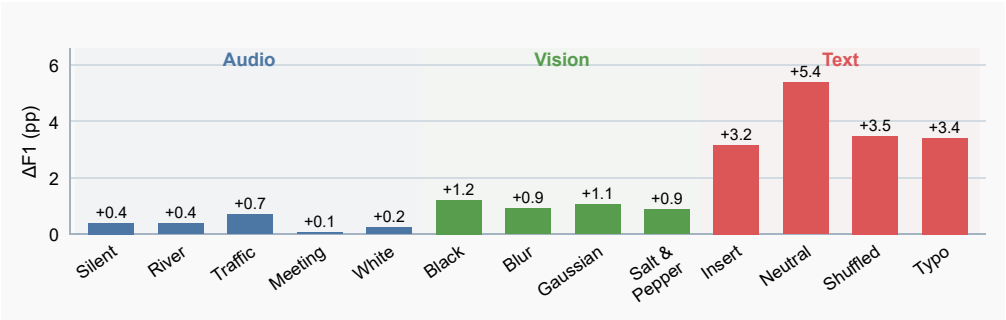


Figure 3: F1 gain under single-modality noise. Bars report $\Delta F_1 = 100(F_1^{\text{Ours}} - F_1^{\text{w/o Denoising}})$ in percentage points (pp), grouped by corrupted modality. All 13 corruption conditions show a positive gain; the underlying values are listed in Table 9.

gains are modest for audio noise (0.07–0.69 points) and larger for visual noise (0.88–1.21 points), while text perturbations produce the strongest gains (3.16–5.39 points).

Paired-modality corruptions. Table 2 summarizes weighted F_1 over all 55 paired-corruption conditions. E-MIND improves each modality-pair average, with the largest gain for text–audio corruption (+7.6 points). The Appendix provides both exhaustive condition-level results and selected cases for interpretation.

Table 2: Paired-corruption weighted F_1 (%) on CMU-MOSI. Means cover all 55 conditions; n counts conditions, not runs. ΔF_1 (pp) is computed before rounding.

Noise pair	n	w/o Denoising	E-MIND	ΔF_1
Text + Audio	20	49.7	**57.3**	+7.6
Text + Vision	16	46.1	**47.2**	+1.1
Audio + Vision	19	84.2	**85.1**	+0.8

4.4 RECOVERY ANALYSIS

We separate a descriptive comparison with imputation references from a controlled ablation of the recovery losses. Their results are shown side by side in Tables 3 and 4.

Table 3: Imputation reference comparison (F_1/ACC_7, %). Full E-MIND values average MR 0.1–0.7.

Dataset	Method	$F_1 \uparrow$	$ACC_7 \uparrow$
CMU-MOSI	w/o Imputation	74.4	36.0
	E-MIND	**80.4**	**38.4**
CMU-MOSEI	w/o Imputation	72.4	47.9
	E-MIND	**79.7**	**51.3**

Table 4: Recovery-loss ablation on CMU-MOSI at nominal MR 0.7 (%).

Variant	$F_1 \uparrow$	$ACC_7 \uparrow$
Flow-only	60.7	25.1
Contrastive-only	70.3	31.0
Full E-MIND	**71.4**	**31.8**

Imputation reference comparison. Table 3 compares full E-MIND with w/o Imputation references. The CMU-MOSI reference is source-reported. On CMU-MOSEI, the reported full E-MIND averages exceed the no-imputation reference by 7.3 percentage points in F_1 and 3.4 points in ACC_7. The no-imputation reference uses one pretrained model on 4,659 CMU-MOSEI test samples with mask seed 1111. Its sample-level masks follow the IMDer missing-one/two proportions and retain at least one modality, so nominal MR 0.7 corresponds to an actual missing fraction of $2/3$. This reference is interpreted descriptively rather than as a controlled ablation of the recovery objective.

Effect of the recovery losses. Table 4 compares the recovery-loss variants on CMU-MOSI at nominal MR 0.7. Full E-MIND achieves the highest scores, exceeding Contrastive-only by 1.1 percentage points in F_1 and 0.8 points in ACC_7.

4.5 EXTENSION TO AVMNIST

To examine conditional energy recovery beyond multimodal sentiment analysis, we evaluate an endpoint-supervised variant on AVMNIST while keeping the complete-modality encoders and classifier frozen. Table 5 compares Energy with two non-conditional feature-filling controls when every test sample has exactly one missing modality. For mixed missingness, half of the samples lack audio and half lack image; no sample loses both modalities. Energy exceeds the stronger filling control by 0.4 points for audio missing, 7.0 points for image missing, and 4.3 points for mixed missingness. Full missing-rate curves and mean $\pm$ standard deviation results appear in Appendix B.

Table 5: AVMNIST accuracy (%; three-seed mean). Mixed: half audio-missing, half image-missing. Δ (pp): gain over the stronger filling baseline.

Missing	Energy	Zero-fill	Mean-fill	Δ
Audio	**64.6**	63.1	64.2	+0.4
Image	**38.4**	31.4	24.8	+7.0
Mixed	**51.4**	47.1	44.3	+4.3

5 DISCUSSION AND CONCLUSION

We introduced E-MIND, a conditional energy framework that jointly learns relative feature energies and recovery directions. By using the remaining clean modalities as context, the learned modality-specific fields support both imputation from Gaussian noise and denoising from corrupted observations without separate noise-specific training. This connects generation and refinement through a shared learned mechanism, with recovered and retained features jointly supporting sentiment prediction. Experiments demonstrate benefits under modality missingness and corruption, while an extension to AVMNIST shows applicability beyond sentiment analysis. Together, these results support conditional energy modeling as a common approach to recovering useful representations from incomplete or unreliable inputs, highlighting its potential for multimodal prediction across varied input conditions.

Our evaluation focuses on selected tasks and degradation settings. Future work will explore broader applications and further assess the framework's robustness and computational efficiency.

AI USE STATEMENT

In this work, generative AI tools were used to assist with preliminary literature search and review, collaborative manuscript drafting and revision, implementation of the E-Mind code, and figure preparation. The authors reviewed all AI-assisted contributions and take responsibility for the final manuscript, code, and figures.

REPRODUCIBILITY STATEMENT

The conditional recovery equations and their gradient routes appear in Section 3; Algorithms 2 and 1 summarize training and inference, while Section 3.5 gives the training objectives and metric conventions. Section 4 specifies the benchmark splits, feature inputs, missingness protocol, and corruption settings; Appendix D provides condition-specific corruption results, and Appendix B gives the AVMNIST protocol and per-seed summary statistics.

REFERENCES

Guillaume Alain and Yoshua Bengio. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research*, 15(1):3563–3593, 2014.

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pp. 1247–1255. PMLR, 2013.

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. *Advances in Neural Information Processing Systems*, 38:8583–8609, 2026.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in neural information processing systems*, 32, 2019.

Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Improved contrastive divergence training of energy based models. *arXiv preprint arXiv:2012.01316*, 2020.

Alexi Gladstone, Ganesh Nanduru, Md Mofijul Islam, Peixuan Han, Hyeonjeong Ha, Aman Chadha, Yilun Du, Heng Ji, Jundong Li, and Tariq Iqbal. Energy-based transformers are scalable learners and thinkers. In *International Conference on Learning Representations*, volume 2026, pp. 139312–139355, 2026.

Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. *arXiv preprint arXiv:1912.03263*, 2019.

Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 1726–1736, 2024.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 9180–9192, 2021.

Wei Han, Hui Chen, Min-Yen Kan, and Soujanya Poria. Mm-align: Learning optimal transport-based alignment dynamics for fast and accurate inference on missing modality sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10498–10511, 2022.

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.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 7837–7851, 2022.

Aapo Hyvärinen and Peter Dayan. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(4):695, 2005.

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.

Mingcheng Li, Dingkang Yang, Yang Liu, Shunli Wang, Jiawei Chen, Shuaibing Wang, Jinjie Wei, Yue Jiang, Qingyao Xu, Xiaolu Hou, et al. Toward robust incomplete multimodal sentiment analysis via hierarchical representation learning. *Advances in Neural Information Processing Systems*, 37:28515–28536, 2024b.

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 *EMNLP (Findings)*, pp. 11252–11267, 2025.

Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

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 (volume 1: Long papers)*, pp. 2247–2256, 2018.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11451–11461. IEEE, 2022.

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.

Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38 (8):1692–1706, 2015.

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.

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, 2020.

Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 47(2):742–754, 2024.

Yuge Shi, Brooks Paige, Philip Torr, et al. Variational mixture-of-experts autoencoders for multi-modal deep generative models. *Advances in neural information processing systems*, 32, 2019.

Yuntao Shou, Jun Yao, Tao Meng, Wei Ai, Cen Chen, and Keqin Li. Gsdnet: Revisiting incomplete multimodality-diffusion emotion recognition from the perspective of graph spectrum. In *IJCAI*, pp. 6182–6190, 2025.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

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, 2023.

Thomas M Sutter, Imant Daunhawer, and Julia E Vogt. Generalized multimodal elbo. *arXiv preprint arXiv:2105.02470*, 2021.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

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, 2019.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *European Conference on Computer Vision*, pp. 575–589. Springer, 2018.

Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. *Advances in Neural Information Processing Systems*, 36:17117–17128, 2023.

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *Advances in neural information processing systems*, 31, 2018.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 438–446, 2024.

Jinsung Yoon, James Jordon, and Mihaela Schaar. Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pp. 5689–5698. PMLR, 2018.

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.

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.

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

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/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.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 2236–2246, 2018b.

Jiandian Zeng, Jiantao Zhou, and Tianyi Liu. Mitigating inconsistencies in multimodal sentiment analysis under uncertain missing modalities. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 2924–2934, 2022.

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. *Advances in Neural Information Processing Systems*, 37:55943–55974, 2024.

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 and the 11th international joint conference on natural language processing (volume 1: Long papers)*, pp. 2608–2618, 2021.

Yan Zhuang, Minhao Liu, Yanru Zhang, Jiawen Deng, and Fuji Ren. Tmdc: A two-stage modality denoising and complementation framework for multimodal sentiment analysis with missing and noisy modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 2281–2289, 2026.

A TRAINING AND RECOVERY ALGORITHMS

Algorithm 1 Core recovery under separate missingness and corruption settings

Require: Projected features, available set O, designated corrupted set C, energies V_{θ_m}; either $C = \varnothing$ or $O = \mathcal{M}$

if $C = \varnothing$ **then** ▷ Missingness or complete clean inputs
  Form fixed condition c_O from observed features
  for $m \in \mathcal{M} \setminus O$ **do**
    Initialize $z_m^{(0)} \sim \mathcal{N}(0, I)$; recover $\widehat{z}_m$ using Eq. 3 with c_O
  end for
else ▷ Corruption with all modalities observed
  Form fixed clean condition $c_{O \setminus C}$, or use no condition if the set is empty
  for $m \in C$ **do**
    Initialize $z_m^{(0)} = \widetilde{z}_m$; recover $\widehat{z}_m$ using Eq. 5
  end for
end if
Assemble recovered and retained features and predict $\widehat{y}$ using Eq. 1
return Sentiment prediction $\widehat{y}$

Algorithm 2 Learning conditional energies for imputation and denoising

Require: Fully observed training utterances, labels, missingness schedule

for each training batch **do**
  Select an available set according to the configured protocol
  for each hidden modality m **do**
    Draw independent noise and interpolation time; compute Eq. 7
    In the main phase, initialize split data/noise negative chains
    Compute Eq. 8, detaching the negative trajectories
    Recover the hidden representation with detached sampling
  end for
  Compute sentiment loss and any enabled post-sampling penalties
  Update parameters using the weighted loss and configured optimizer
end for
return Learned energies $\{V_{\theta_m}\}_{m \in \mathcal{M}}$ and predictor f_ϕ for Algorithm 1

B AVMNIST: FEATURE COMPLETION BEYOND SENTIMENT ANALYSIS

Data and evaluation protocol. AVMNIST pairs handwritten digit images with spoken-digit spectrograms for ten-class classification (Vielzeuf et al., 2018). We use 55,000 training, 5,000 validation, and 10,000 test examples, with 28×28 images and 112×112 audio spectrograms. Each modality encoder produces a 32-dimensional feature. For each seed (1111, 1112, and 1113), a classifier is first trained with complete modalities; its encoders and fusion classifier are then frozen, and only the conditional energy models are trained for completion. All three completion methods use the same frozen checkpoint and missingness masks within each seed.

Here, missing rate denotes the fraction of test samples with one unavailable modality, rather than the fraction of modality slots removed in Equation 10. In audio-missing and image-missing evaluations, only the named modality is removed from that fraction of samples. Mixed missingness assigns half of incomplete samples to each missing modality; both modalities are never removed from the same sample. At a rate of 100%, every test example therefore retains exactly one modality. Recovered features are reused across missing rates within each run.

Filling baselines. Zero-fill replaces the missing encoder-output feature with an all-zero vector. Mean-fill replaces it with the corresponding modality's per-dimension training-set feature mean. These operations are defined in the original encoder feature space: zero in the standardized energy-

model space corresponds to mean-fill after inverse normalization. Neither baseline conditions the replacement on the observed modality.

Endpoint-supervised energy variant. The AVMNIST results use an extension of the flow-regression and contrastive objectives with endpoint reconstruction and classification supervision:

$$\mathcal{L}_{\mathrm{AV}} = \mathcal{L}_{\mathrm{FM}} + 0.001\mathcal{L}_{\mathrm{CD}} + \|\hat{z}_m - z_m\|_2^2 + \mathrm{CE}(C(z_o, \hat{z}_m), y)\,, \tag{11}$$

where the reconstruction term uses standardized features and the frozen classifier C receives recovered features after inverse normalization. The observed modality is denoted by o and the missing modality by m. Starting from the original energy checkpoint, we fine-tune each modality's energy model for 20 epochs using AdamW (learning rate 3×10^{-4}, batch size 128). Recovery starts from Gaussian noise and uses five energy-gradient flow updates (step size 0.2), followed by five Langevin updates (step size 0.01, noise scale 0.01). Training and evaluation use the same recovery schedule. Endpoint losses backpropagate through all recovery updates to the energy parameters; the encoders and classifier remain frozen. Checkpoints are selected by validation accuracy under the corresponding missing modality. Thus, these results evaluate the endpoint-supervised extension, rather than the flow-plus-contrastive objective alone.

Full results. Figure 4 and Tables 6–8 report test accuracy as mean ± sample standard deviation over three seeds at all eleven missing rates. Differences are calculated from unrounded means and expressed in percentage points. The aggregate accuracy gap grows as more examples require completion; this does not imply that recovery quality per incomplete example improves with missing rate.

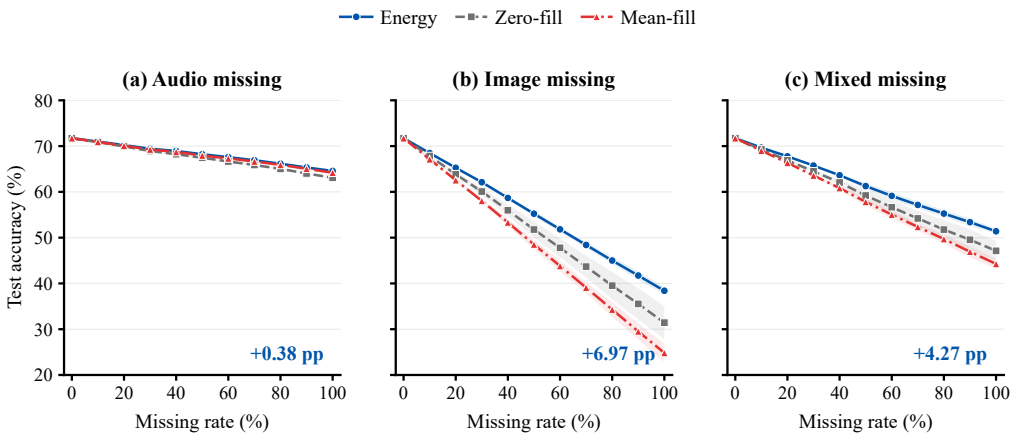


Figure 4: Full AVMNIST missing-rate curves for Audio missing, Image missing, and Mixed missing. Energy, Zero-fill, and Mean-fill use the same frozen classifier and missingness masks within each seed. Shading denotes the sample standard deviation across three seeds; annotations give Energy minus the stronger filling control at 100% incompleteness.

C Recovery Analysis

These statements explain the inference objective under explicit assumptions. They do not prove that the trained model satisfies those assumptions, reaches equilibrium, or improves a downstream metric.

C.1 Analysis of the Recovery Objective

Autonomous flow regression as a projection. Consider uniform weighting in Eq. 7. Let $Z = (1-t)\xi + tz_m$, $U = z_m - \xi$, and let Q denote the random conditioning representation. Expectations

Table 6: AVMNIST: Audio missing. Test accuracy (%), mean ± sample standard deviation across three seeds. Δ is Energy minus Zero-fill (pp).

Missing rate (%)	Energy	Zero-fill	Mean-fill	Δ
0	71.7 ± 0.4	71.7 ± 0.4	71.7 ± 0.4	$+0.0$
10	71.0 ± 0.2	70.9 ± 0.2	70.9 ± 0.2	$+0.1$
20	70.2 ± 0.1	69.9 ± 0.0	70.1 ± 0.1	$+0.3$
30	69.4 ± 0.2	68.9 ± 0.3	69.3 ± 0.2	$+0.5$
40	68.9 ± 0.2	68.2 ± 0.3	68.7 ± 0.2	$+0.7$
50	68.2 ± 0.1	67.5 ± 0.3	68.0 ± 0.2	$+0.8$
60	67.6 ± 0.1	66.6 ± 0.3	67.3 ± 0.2	$+1.0$
70	66.9 ± 0.2	65.9 ± 0.3	66.6 ± 0.2	$+1.0$
80	66.1 ± 0.2	65.0 ± 0.2	65.9 ± 0.2	$+1.1$
90	65.3 ± 0.3	64.0 ± 0.3	65.1 ± 0.3	$+1.3$
100	64.6 ± 0.3	63.1 ± 0.2	64.2 ± 0.3	$+1.5$

Table 7: AVMNIST: Image missing. Test accuracy (%), mean ± sample standard deviation across three seeds. Δ is Energy minus Zero-fill (pp).

Missing rate (%)	Energy	Zero-fill	Mean-fill	Δ
0	71.7 ± 0.4	71.7 ± 0.4	71.7 ± 0.4	$+0.0$
10	68.5 ± 0.2	67.8 ± 0.3	67.1 ± 0.2	$+0.7$
20	65.3 ± 0.3	63.8 ± 0.6	62.6 ± 0.3	$+1.4$
30	62.1 ± 0.3	60.1 ± 0.8	58.0 ± 0.4	$+2.1$
40	58.7 ± 0.3	56.0 ± 1.2	53.3 ± 0.8	$+2.7$
50	55.2 ± 0.6	51.8 ± 1.8	48.5 ± 1.0	$+3.5$
60	51.8 ± 0.6	47.8 ± 2.1	43.8 ± 1.2	$+4.0$
70	48.4 ± 0.6	43.7 ± 2.4	39.1 ± 1.2	$+4.7$
80	45.0 ± 0.8	39.5 ± 3.0	34.3 ± 1.6	$+5.5$
90	41.7 ± 1.0	35.5 ± 3.4	29.5 ± 1.8	$+6.2$
100	38.4 ± 0.8	31.4 ± 3.7	24.8 ± 2.0	$+7.0$

Table 8: AVMNIST: Mixed missing. Test accuracy (%), mean ± sample standard deviation across three seeds. Δ is Energy minus Zero-fill (pp).

Missing rate (%)	Energy	Zero-fill	Mean-fill	Δ
0	71.7 ± 0.4	71.7 ± 0.4	71.7 ± 0.4	$+0.0$
10	69.7 ± 0.2	69.3 ± 0.3	69.0 ± 0.2	$+0.4$
20	67.8 ± 0.2	66.9 ± 0.4	66.3 ± 0.2	$+0.9$
30	65.7 ± 0.1	64.5 ± 0.4	63.6 ± 0.2	$+1.3$
40	63.6 ± 0.2	62.1 ± 0.7	60.8 ± 0.5	$+1.6$
50	61.3 ± 0.6	59.2 ± 1.2	57.8 ± 0.8	$+2.1$
60	59.1 ± 0.8	56.6 ± 1.6	55.0 ± 0.9	$+2.5$
70	57.2 ± 0.8	54.2 ± 1.6	52.4 ± 0.8	$+3.0$
80	55.3 ± 0.8	51.8 ± 2.1	49.7 ± 1.0	$+3.5$
90	53.4 ± 0.8	49.5 ± 2.3	46.9 ± 1.2	$+3.8$
100	51.4 ± 0.5	47.1 ± 2.2	44.3 ± 1.1	$+4.3$

are over training utterances, noise, and interpolation time. For any square-integrable field $v(Z, Q)$, conditional expectation gives

$$\mathbb{E}\|v(Z,Q)-U\|^2 = \mathbb{E}\|v(Z,Q)-\bar{u}(Z,Q)\|^2+\mathbb{E}\|U-\bar{u}(Z,Q)\|^2, \quad \bar{u}(Z,Q) = \mathbb{E}[U \mid Z,Q]. \tag{12}$$

The cross term vanishes because $\mathbb{E}[U - \bar{u} \mid Z, Q] = 0$. An unrestricted autonomous regressor therefore estimates the conditional mean with interpolation time marginalized out. Parameterizing v as a negative potential gradient further constrains the admissible fields. The flow loss provides a regression signal within this class, but does not establish exact time-conditioned transport to the data distribution. Label-dependent weighting changes the effective regression distribution and requires separate evaluation.

Relation to the three theory references. Energy Matching (Balcerak et al., 2026) derives a transport-to-equilibrium construction and uses explicit transport coupling in its training formulation. Our independent pairing and time support are implementation choices, so its exact transport interpretation is not asserted here. Energy-Based Transformers (Gladstone et al., 2026) optimize candidate predictions and train through their optimization trajectories, unlike our detached feature recovery. Equilibrium Matching (Wang & Du, 2025) uses endpoint-vanishing targets; Eq. 6 does not. The analysis above characterizes autonomous regression under explicit assumptions.

D Detailed corruption results

All triplets in this section report percentages in the order $ACC_2/F_1/ACC_7$, rounded to one decimal place; delta columns give percentage-point changes, computed before rounding and displayed to one decimal place. Emphasis inherited from the source should be read alongside the per-metric values, since a gain in binary accuracy need not imply gains in all metrics.

D.1 Single-modality corruption

Table 9 contains 13 settings. Denoising improves binary accuracy for the listed audio and visual corruptions, but reduces it for Insert and Shuffled text.

Table 9: Detailed robustness evaluation under **single-modality** noise conditions. Metrics are reported as ACC_2 / F_1 / ACC_7 (%).

Modality	Noise Type	w/o Denoising (Baseline)	Ours (E-MIND)	ΔF_1
Audio	Silent	82.4 / 84.4 / 46.3	**82.9 / 84.8 / 46.6**	↑ 0.4 pp
	River Noise	82.9 / 84.9 / 45.2	**83.2 / 85.2 / 47.3**	↑ 0.4 pp
	Traffic Noise	82.6 / 84.6 / 45.5	**83.2 / 85.2 / 46.0**	↑ 0.7 pp
	Meeting Noise	82.9 / 84.9 / 46.1	**83.0 / 84.9 / 46.7**	↑ 0.1 pp
	White Noise	82.7 / 84.7 / 45.7	**82.9 / 84.9 / 46.7**	↑ 0.2 pp
Vision	Black	82.6 / 84.4 / 44.1	**83.8 / 85.6 / 45.7**	↑ 1.2 pp
	Blur	82.7 / 84.6 / **47.0**	**83.6 / 85.5** / 45.7	↑ 0.9 pp
	Gaussian	82.7 / 84.4 / 46.0	**83.6 / 85.5 / 46.1**	↑ 1.1 pp
	Salt & Pepper	82.7 / 84.6 / **47.4**	**83.6 / 85.5** / 46.3	↑ 0.9 pp
Text	Insert	**55.8** / 52.8 / **19.6**	55.2 / **55.9** / 19.0	↑ 3.2 pp
	Neutral	54.2 / 50.4 / **19.9**	**55.1 / 55.8** / 19.2	↑ 5.4 pp
	Shuffled	**55.6** / 52.4 / **21.4**	55.2 / **55.9** / 19.3	↑ 3.5 pp
	Typo	52.6 / 52.7 / 17.4	**55.3 / 56.1 / 19.3**	↑ 3.4 pp

D.2 Paired corruption

The following tables contain 19 audio–vision, 20 text–audio, and 16 text–vision combinations. They include improvements and regressions. Table 10 gives selected examples with substantial ACC_2 gains; the exhaustive comparisons below remain the basis for the aggregate results in Table 2.

Table 10: Selected paired-corruption cases with substantial ACC_2 gains on CMU-MOSI. Entries report $ACC_2/F_1/ACC_7$ (%); ΔACC_2 is in percentage points (pp). Deltas are computed before rounding.

Modality	Noise Combination	w/o Denoising (Baseline)	Ours (E-MIND)	ΔACC_2
A + V	Silent + Blur	81.8 / 83.9 / 45.5	83.6 / 85.6 / 46.1	↑ **1.8 pp**
	White Noise + S&P	81.7 / 83.8 / 47.0	83.5 / 85.4 / 46.3	↑ **1.8 pp**
	Meeting Noise + Black	82.1 / 84.1 / 43.8	83.3 / 85.3 / 45.7	↑ **1.2 pp**
T + A	Shuffled + Silent	53.0 / 48.3 / 19.8	57.0 / 57.8 / 21.1	↑ **4.0 pp**
	Neutral + Silent	52.9 / 47.0 / 19.6	56.7 / 57.5 / 21.2	↑ **3.8 pp**
	Insert + River	52.3 / 48.3 / 17.7	55.9 / 56.7 / 20.8	↑ **3.7 pp**
T + V	Shuffled + Black	45.7 / 42.6 / 15.4	53.9 / 47.8 / 15.4	↑ **8.2 pp**
	Shuffled + Blur	46.0 / 42.8 / 15.8	52.3 / 45.4 / 16.3	↑ **6.3 pp**
	Neutral + Blur	48.5 / 47.1 / 14.9	54.5 / 43.0 / 17.6	↑ **6.0 pp**

E METRIC-SPECIFIC TRADE-OFFS AND INTERPRETATION

A row-wise comparison with IMDer uses the displayed settings to separate binary decisions from sentiment-intensity recovery. Across the eight displayed MOSI settings, E-MIND has higher F_1 in all eight, but higher ACC_7 in only two. On MOSEI, it has higher F_1 in six settings and higher ACC_7 in seven. With MOMKE and TMDC included, E-MIND still leads MOSI F_1 in all eight settings and MOSEI ACC_7 in seven, whereas TMDC leads MOSEI F_1 at MR 0.2–0.7. These counts describe the displayed numbers; they are neither additional trials nor estimates of statistical confidence.

The difference between these patterns matters for the recovery objective. Preserving the sign of a sentiment prediction is a weaker requirement than preserving its intensity bin. A feature change can move a prediction to the correct side of zero while leaving its magnitude inaccurate. The MOSI results are compatible with that distinction: improvements in binary classification coexist with weaker seven-class performance at several missingness levels. They do not establish that recovered features are closer to their hidden targets in every metric or that the learned conditional distribution is calibrated. Reconstruction error and conditional sample diagnostics would be needed to test those stronger interpretations directly.

The corruption comparisons expose a related distinction between observation recovery and decision recovery. Under visual blur, the single-modality table reports higher binary accuracy after refinement, but lower seven-class accuracy. Under inserted text, binary accuracy decreases while weighted F_1 increases. We therefore interpret the method through the full metric triplet and the corruption condition, rather than reducing recovery quality to one score. Together, these comparisons locate the present empirical support at the level of task-specific decision changes; the learned energy field provides a mechanism to investigate, not a substitute for measuring feature fidelity.

Energy values also require care when used as diagnostics. Adding a modality-specific constant to a potential leaves its recovery gradient unchanged and cancels in the contrastive energy difference. The core objectives therefore do not fix a common absolute energy scale across modalities. A threshold on one modality's potential cannot automatically be transferred to another, even if both support useful recovery. Assessing corruption detection consequently requires a specified scoring rule, a fixed score orientation, and calibration on held-out data. Recovery accuracy alone does not establish these properties.

Table 11: Exhaustive robustness evaluation under simultaneous **Audio and Vision** noise. Metrics are reported as ACC_2 / F_1 / ACC_7 (%).

Noise Combination	**w/o Denoising (Baseline)**	**Ours (E-MIND)**	ΔF_1
Audio: Meeting + Vision: Black	82.1 / 84.1 / 43.8	**83.3 / 85.3 / 45.7**	↑ 1.2 pp
Audio: Meeting + Vision: Blur	82.3 / 84.4 / 44.2	**83.3 / 85.3 / 45.5**	↑ 0.9 pp
Audio: Meeting + Vision: Salt & Pepper	82.3 / 84.4 / 44.5	**83.0 / 85.0 / 46.1**	↑ 0.6 pp
Audio: River + Vision: Black	82.4 / 84.2 / 43.2	**83.0 / 85.0 / 45.1**	↑ 0.7 pp
Audio: River + Vision: Blur	82.9 / 84.7 / 44.8	**83.5 / 85.4 / 46.1**	↑ 0.7 pp
Audio: River + Vision: Gaussian	82.3 / 84.1 / 45.5	**82.6 / 84.4 / 45.0**	↑ 0.3 pp
Audio: River + Vision: Salt & Pepper	82.6 / 84.3 / 46.3	**83.5 / 85.4 / 45.4**	↑ 1.2 pp
Audio: Silent + Vision: Black	82.7 / 84.7 / 43.9	**83.5 / 85.4 / 46.0**	↑ 0.7 pp
Audio: Silent + Vision: Blur	81.8 / 83.9 / 45.5	**83.6 / 85.6 / 46.1**	↑ 1.6 pp
Audio: Silent + Vision: Gaussian	82.0 / 83.9 / 43.1	**82.0 / 83.7 / 44.7**	↓ 0.3 pp
Audio: Silent + Vision: Salt & Pepper	82.3 / 84.1 / 45.0	**83.5 / 85.4 / 46.1**	↑ 1.3 pp
Audio: Traffic + Vision: Black	82.1 / 84.1 / 44.4	**83.3 / 85.3 / 45.2**	↑ 1.2 pp
Audio: Traffic + Vision: Blur	82.6 / 84.4 / 46.9	**83.2 / 85.1 / 45.2**	↑ 0.7 pp
Audio: Traffic + Vision: Gaussian	82.0 / 83.9 / 45.2	**82.4 / 84.1 / 44.1**	↑ 0.2 pp
Audio: Traffic + Vision: Salt & Pepper	82.0 / 83.8 / 47.9	**83.3 / 85.3 / 45.7**	↑ 1.5 pp
Audio: White + Vision: Black	82.6 / 84.4 / 44.8	**83.5 / 85.4 / 45.0**	↑ 1.0 pp
Audio: White + Vision: Blur	82.9 / 84.7 / 45.4	**83.5 / 85.4 / 45.5**	↑ 0.7 pp
Audio: White + Vision: Gaussian	82.3 / 84.2 / 45.4	**82.4 / 84.3 / 44.1**	↑ 0.0 pp
Audio: White + Vision: Salt & Pepper	81.7 / 83.8 / 47.0	**83.5 / 85.4 / 46.3**	↑ 1.6 pp

Table 12: Exhaustive robustness evaluation under simultaneous **Text and Audio** noise.

Noise Combination	**w/o Denoising (Baseline)**	**Ours (E-MIND)**	ΔF_1
Text: Insert + Audio: River	52.3 / 48.3 / 17.7	**55.9 / 56.7 / 20.8**	↑ 8.4 pp
Text: Insert + Audio: Silent	52.7 / 48.1 / 20.4	**56.4 / 57.2 / 21.5**	↑ 9.2 pp
Text: Insert + Audio: White	52.3 / 48.9 / 20.1	**56.2 / 57.1 / 20.4**	↑ 8.2 pp
Text: Insert + Audio: Meeting	52.1 / 49.4 / 19.3	**56.2 / 57.1 / 21.1**	↑ 7.7 pp
Text: Insert + Audio: Traffic	53.2 / 51.0 / 18.3	**57.1 / 58.0 / 21.1**	↑ 7.0 pp
Text: Neutral + Audio: River	53.0 / 48.7 / 20.2	**56.1 / 56.9 / 21.5**	↑ 8.2 pp
Text: Neutral + Audio: Silent	52.9 / 47.0 / 19.6	**56.7 / 57.5 / 21.2**	↑ 10.5 pp
Text: Neutral + Audio: Traffic	53.7 / 49.9 / 19.8	**56.8 / 57.7 / 21.2**	↑ 7.8 pp
Text: Neutral + Audio: Meeting	54.6 / 50.0 / 20.8	**55.9 / 56.6 / 21.1**	↑ 6.6 pp
Text: Neutral + Audio: White	53.0 / 46.3 / 20.5	**55.8 / 56.4 / 21.1**	↑ 10.1 pp
Text: Shuffled + Audio: River	53.4 / 48.6 / 19.9	**56.7 / 57.5 / 20.9**	↑ 9.0 pp
Text: Shuffled + Audio: Silent	53.0 / 48.3 / 19.8	**57.0 / 57.8 / 21.1**	↑ 9.5 pp
Text: Shuffled + Audio: Traffic	54.5 / 51.3 / 20.9	**56.7 / 57.5 / 20.5**	↑ 6.3 pp
Text: Shuffled + Audio: Meeting	48.9 / 46.4 / 18.9	**56.7 / 57.5 / 20.6**	↑ 11.2 pp
Text: Shuffled + Audio: White	53.0 / 50.0 / 19.2	**56.2 / 57.1 / 21.1**	↑ 7.0 pp
Text: Typo + Audio: River	52.4 / 52.1 / 19.9	**56.1 / 56.9 / 20.5**	↑ 4.8 pp
Text: Typo + Audio: Silent	56.1 / 55.0 / 17.3	**57.3 / 58.1 / 21.1**	↑ 3.1 pp
Text: Typo + Audio: White	50.7 / 51.2 / 21.4	**56.5 / 57.4 / 20.4**	↑ 6.2 pp
Text: Typo + Audio: Meeting	52.4 / 53.1 / 19.2	**57.0 / 57.8 / 20.8**	↑ 4.7 pp
Text: Typo + Audio: Traffic	50.8 / 51.4 / 17.1	**57.0 / 57.8 / 21.2**	↑ 6.5 pp

Table 13: Exhaustive robustness evaluation under simultaneous **Text and Vision** noise.

Noise Combination	**w/o Denoising (Baseline)**	**Ours (E-MIND)**	ΔACC_2
Text: Insert + Vision: Black	49.2 / 48.5 / 16.3	**54.6 / 47.8 / 16.3**	↑ 5.4 pp
Text: Insert + Vision: Blur	52.4 / 47.6 / 18.3	**55.9 / 49.7 / 15.7**	↑ 3.5 pp
Text: Insert + Vision: Gaussian	46.9 / 44.6 / 15.5	**48.0 / 46.4 / 14.6**	↑ 1.2 pp
Text: Insert + Vision: Salt & Pepper	51.2 / 46.1 / 19.8	**47.9 / 47.1 / 15.1**	↓ 3.4 pp
Text: Neutral + Vision: Black	51.5 / 48.1 / 18.5	**53.9 / 42.4 / 18.2**	↑ 2.3 pp
Text: Neutral + Vision: Blur	48.5 / 47.1 / 14.9	**54.5 / 43.0 / 17.6**	↑ 6.0 pp
Text: Neutral + Vision: Gaussian	54.0 / 51.3 / 20.5	**54.5 / 53.3 / 21.8**	↑ 0.4 pp
Text: Neutral + Vision: Salt & Pepper	53.0 / 47.8 / 19.9	**51.8 / 49.8 / 18.6**	↓ 1.2 pp
Text: Shuffled + Vision: Black	45.7 / 42.6 / 15.4	**53.9 / 47.8 / 15.4**	↑ 8.2 pp
Text: Shuffled + Vision: Blur	46.0 / 42.8 / 15.8	**52.3 / 45.4 / 16.3**	↑ 6.3 pp
Text: Shuffled + Vision: Gaussian	50.4 / 49.3 / 20.5	**50.1 / 50.8 / 21.8**	↓ 0.3 pp
Text: Shuffled + Vision: Salt & Pepper	52.1 / 46.6 / 19.3	**50.1 / 49.8 / 18.0**	↓ 2.1 pp
Text: Typo + Vision: Black	49.8 / 49.5 / 16.1	**50.7 / 50.5 / 16.3**	↑ 0.9 pp
Text: Typo + Vision: Blur	46.3 / 37.2 / 15.2	**48.6 / 47.5 / 16.4**	↑ 2.3 pp
Text: Typo + Vision: Gaussian	49.8 / 50.2 / 17.0	**46.6 / 42.8 / 18.3**	↓ 3.2 pp
Text: Typo + Vision: Salt & Pepper	46.4 / 38.7 / 15.1	**47.4 / 40.9 / 14.4**	↑ 1.0 pp