# CHANNEL-AWARE TURBO MULTIMODAL FUSION: LEARNING UNDER NOISY AND MISSING MODALITIES

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Multimodal systems in the wild must operate when modalities are corrupted by noise or entirely missing. Most existing methods address the two failure modes separately and fuse modalities with attention that is blind to signal quality. We draw an analogy between multimodal fusion and multi-channel communication, and propose **Channel-Aware Turbo Multimodal Fusion** (CAT-MF), which treats noise and missingness uniformly as channel impairments. CAT-MF has three components: (1) **channel state estimation**, which scores every modality by an existence probability, a spectral proxy for the inverse signal-to-noise ratio, and a learned damage-type embedding, all obtained without clean reference signals; (2) **iterative joint decoding**, a Turbo-inspired procedure that exchanges soft cross-modal information over a small number of rounds and accumulates it in a persistent state; and (3) **channel-modulated attention**, which gates cross-modal transfer by the estimated states, zeroing missing channels, down-weighting noisy ones, and suppressing transfer between channels sharing correlated damage. We prove that the decoding iterations converge absolutely and linearly to a limit once the steps are summable, that the accumulated extrinsic evidence yields non-negative information gain per round bounded below by the donor modalities' quality, and that the reconstruction error decays with the availability of high-quality correlated modalities. CAT-MF exceeds competitive baselines on multiple datasets and is markedly more robust under noisy and missing modalities.

## 1 INTRODUCTION

Multimodal systems integrate complementary signals such as language, vision, and audio to interpret human behavior and perceive the physical world (Baltrušaitis et al., 2019; Liang et al., 2022; Zhuang et al., 2026). Real deployment, however, routinely violates the clean-and-complete assumption under which most fusion models are trained: acoustic tracks pick up background noise, visual streams suffer occlusion or blur, transcripts inherit ASR errors, and entire modalities vanish because of sensor failure, privacy restrictions, or packet loss (Zhang et al., 2023b). Two observations make this difficult. First, corruption and missingness are not isolated events but co-occur and interact: a noisy acoustic channel paired with an occluded visual stream is harder to handle than either defect alone. However, the literature has largely built two separate toolboxes, relying on denoising, adversarial augmentation, and robust training for the former and on imputation or modality dropout for the latter (Zhang et al., 2024b; Wu et al., 2026). Second, the weakness lies inside the fusion operator itself. Since Transformers (Vaswani et al., 2017) and their multimodal variants (Tsai et al., 2019; Zhang et al., 2023a) became the dominant fusion mechanism, attention weights are determined based on feature similarity, so a loudly corrupted modality can remain highly similar to the query and dominate the joint representation precisely when it should be ignored. Unimodal uncertainty estimates (Gao et al., 2024b; Cho et al., 2025) try to down-weight noisy modalities, yet they act mainly at final aggregation and do not arrest error propagation during cross-modal interaction.

A communication receiver faces the same situation: several noisy, intermittently available channels carry correlated information about one source message, and the receiver must estimate each channel's state, combine channels in proportion to their reliability, and refine its belief iteratively as in Turbo decoding (Berrou et al., 1993). This analogy was explored for classical audio-visual speech fusion (Shivappa et al., 2010) and deep joint source-channel coding (Bourtsoulatze et al., 2019; Hu et al., 2025). These efforts, however, neither learn channel state from data nor target the



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.

cross-modal error propagation that dominates multimodal failure. We instead cast each modality as a channel, the latent semantic content as the source, and the fusion model as the receiver. This view pays off in three ways: (1) noise and missingness become two points on a single impairment axis that admits one estimator and one gating rule; (2) channel state supplies explicit, interpretable reliability signals that are distinct from content relevance in regular similarity-based attention; (3) and the Turbo principle prescribes an iterative soft-information exchange that, under a geometrically decaying schedule, provably converges to a unique limit at a linear rate.

We instantiate this view as **Channel-Aware Turbo Multimodal Fusion** (**CAT-MF**), which reorganizes fusion into *estimating channel state, exchanging soft evidence across modalities with iterative joint decoding, and reading out from reliable channels*. First, a **channel state estimator** evaluates every modality via three complementary signals: (1) an existence probability that marks explicit outages and also catches implicit missingness such as near-silent or fully occluded streams; (2) a deterministic inverse-SNR proxy derived from the normalized high-frequency energy of the observed spectrum; (3) and a damage-type embedding that characterizes *how* a channel is impaired rather than only *how much*. Fusion then proceeds through a small number of **iterative joint decoding rounds** with shared weights, mirroring the exchange of extrinsic information between constituent Turbo decoders: each modality gathers cross-modal evidence, accumulates it in a persistent soft-information state instead of overwriting it, and updates its representation with a bounded residual step, so that expressive depth grows with the number of rounds without stacking distinct attention layers. Crucially, the cross-modal transfer at every round is **channel-modulated** by the estimated states, which gives three behaviors that standard similarity-based attention cannot provide: (1) a missing donor contributes exactly nothing rather than a zero-padded artifact; (2) a noisy donor is exponentially attenuated at every round, containing error propagation during interaction instead of only at the final readout; (3) and two channels whose damage types are aligned exchange less, so that correlated errors cannot corroborate each other. After decoding, a state-weighted fusion feeds the task head, and an auxiliary peer-reconstruction objective restores weak channels from their reliable counterparts.

In short, CAT-MF explicitly separates *reliability* (the channel state) from *relevance* (feature similarity) and enforces reliability throughout the cross-modal interaction rather than after it. The channel-modulated, weight-tied decoder is not merely an effective heuristic: Section 4 gives explicit step-size and channel-state conditions under which the decoder is stable and provably suppresses unreliable channels, replacing the qualitative rule "down-weight bad modalities" with testable guarantees.

The contributions of this paper are summarized as below:

- **A unified communication-theoretic model and architecture.** We formulate noisy and missing modalities as points on one channel-impairment axis and propose CAT-MF, which for the first time combines reference-free per-modality channel state estimation, Turbo-style shared-weight iterative decoding, and channel-modulated attention that suppresses missing, noisy, and correlated-damage donors by distinct mechanisms.
- **Theory Contribution.** We prove that, with a geometrically decaying accumulation schedule, the iterative joint decoding iterations converge absolutely and linearly to a limit as the steps become summable, that the accumulated extrinsic evidence yields non-negative per-round information gain bounded below by donor-channel quality and cross-modal correlation, and that the reconstruction error shrinks as high-quality correlated modalities become available; the channel gate is further shown to attenuate propagated noise variance.
- **Experiments.** On multiple datasets, CAT-MF is competitive with or better than state of the art on clean data and markedly more robust as noise and missing rates grow, while controlled comparisons against stacked cross-attention, ablations, and visualizations isolate the role of each component and confirm the interpretability of the estimated channel states.

## 2 Related Work

**Multimodal fusion and attention.** Early fusion paradigms include tensor factorization (Zadeh et al., 2017), memory networks (Zadeh et al., 2018a), and shared-specific representation decomposition (Hazarika et al., 2020). Recent work explores state-space backbones and parameter-efficient proxy tokens for efficient or incomplete fusion (Reza et al., 2025; Li et al., 2025b). Cross-modal Transformers become dominant, capturing pairwise interactions with latent cross-attention (Tsai

et al., 2019; Mai et al., 2025). In all of these, however, attention weights are learned from the task loss and reflect feature similarity rather than signal fidelity, so a corrupted modality that stays similar to the query retains a large weight. Explicitly supervising or gating attention has been attempted, but they impose absolute importance labels that are hard to estimate without clean references (Mai et al., 2024), or quantify uncertainty only for late aggregation (Gao et al., 2024b; Cho et al., 2025). CAT-MF instead conditions every round of cross-modal transfer on an independently estimated, physically interpretable channel state, and separates this reliability signal from content similarity.

**Learning with noisy and missing modalities.** Missing-modality learning is now a mature subfield (Zhang et al., 2023b; Wu et al., 2026) and a parallel literature studies fusion on noisy, imbalanced, and otherwise low-quality data (Zhang et al., 2024b). Generative methods imagine absent modalities from observed ones, using cascaded translation, masked autoencoders, or, more recently, diffusion priors (Ma et al., 2021; Yuan et al., 2021; Wang et al., 2023). Self-distillation and cross-sample enhancement avoid explicit reconstruction by distilling consistent multimodal representations (Li et al., 2024; Zhuang et al., 2025; Zhang et al., 2024a), and robustness-oriented methods correct dominant modalities or estimate unimodal uncertainty (Gao et al., 2024b;a), while information-bottleneck objectives filter unimodal noise (Mai et al., 2023b; Wu et al., 2025). These lines of work treat corruption and outage separately, and collapse reliability to a single scalar: they cannot tell how a modality is corrupted, nor prevent two channels that fail in the same way from reinforcing one another. CAT-MF handles both impairments with one state triplet and uses the damage type explicitly inside the gate, so that correlated errors are decorrelated rather than aggregated, and CAT-MF controls error propagation during multimodal interaction rather than during prediction.

**Iterative, equilibrium, and communication-inspired models.** Iterative Turbo decoding exchanges extrinsic information between constituent decoders and was adapted to classical audio–visual speech fusion (Berrou et al., 1993; Hagenauer et al., 1996; Shivappa et al., 2010). Deep joint source-channel coding trains neural codecs directly over noisy channels (Bourtsoulatze et al., 2019), and recent multimodal semantic-communication schemes extend this idea by adapting the receiver to estimated channel quality (Hu et al., 2025). Deep equilibrium models compute the fixed point of weight-tied iterative refinement and analyze its stability (Bai et al., 2019). This fixed-point view has been adopted for multimodal fusion (Ni et al., 2023), and recent looped-Transformer work further studies the stability and adaptive halting of such weight-tied iterations (Movahedi et al., 2026). CAT-MF borrows the Turbo principle of accumulated soft evidence and the convergence analysis of equilibrium models, but applies them to semantic cross-modal interaction whose messages are modulated by learned channel state, rather than to signal transmission or generic hidden states.

To our best knowledge, CAT-MF is the first framework to combine (1) learned, reference-free channel state estimation that jointly captures existence, noise level, and damage type, (2) Turbo-style shared-weight iterative decoding with persistent soft-information accumulation, and (3) channel-modulated attention that enforces reliability at every interaction round, and to accompany the design with convergence, information-gain, reconstruction, and noise-suppression guarantees.

## 3 Channel-Aware Turbo Multimodal Fusion

Figure 1 shows the pipeline: feature extraction, channel estimation, iterative joint decoding, and state-weighted fusion. Throughout, $M$ modalities are indexed by $m, n$, the common feature dimension is $d$, and a batch element is suppressed from the notation.

Let $s \in \mathbb{R}^{d_s}$ denote the latent semantic content (e.g., affect). For $M$ modalities, a clean feature $x_m = f_m(s) \in \mathbb{R}^{d_m}$ passes through a degradation channel $\mathcal{C}_m$ to yield the observation

$$y_m = \mathcal{C}_m(x_m), \qquad \mathcal{C}_m : \;\; y_m = x_m + n_m, \quad \text{or} \quad y_m = \varnothing \text{ w.p. } p_m^{\text{miss}} \text{ (outage)}, \tag{1}$$

where $n_m$ is additive channel noise and the two events describe, respectively, a corrupted and a missing modality; both may occur in one mini-batch.

### 3.1 Feature Extraction

Without loss of generality, assuming that the modality features are in the form of sequences or vectors, each modality is first encoded within itself into the common $d$-dimensional space and then

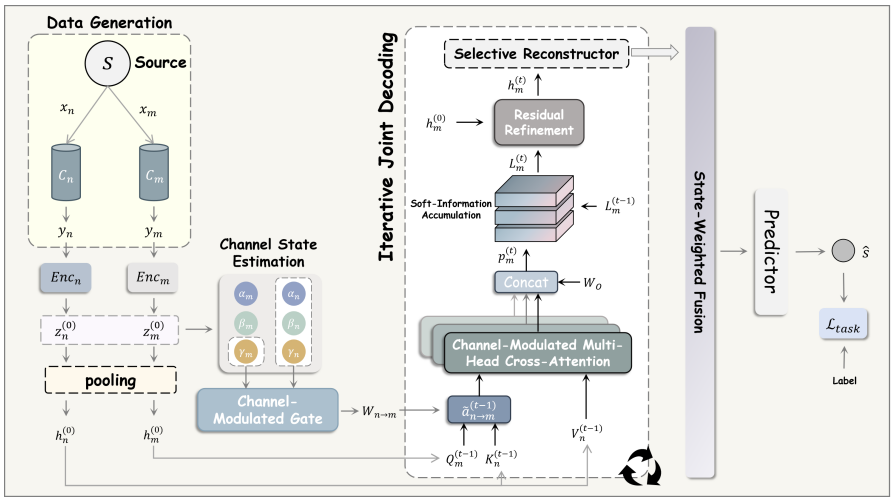



Figure 1: The pipeline of CAT-MF. Two modalities are used for illustration.

pooled to one token. Concretely, a modality-specific front end produces a sequence:

$$z_m^{(0)} = \widetilde{\mathrm{Enc}}_m(y_m) \in \mathbb{R}^{T_m \times d},$$

The final output of the encoder is the temporal mean:

$$h_m^{(0)} = \mathrm{Enc}_m(y_m) = \begin{cases} \frac{1}{T_m}\sum_{t=1}^{T_m} z_{m,t}^{(0)} \in \mathbb{R}^d, & y_m \text{ is a sequence,} \\ z_m^{(0)} \in \mathbb{R}^d, & y_m \text{ has no time axis,} \end{cases} \tag{2}$$

and we stack $H^{(0)} = [h_1^{(0)}; \ldots; h_M^{(0)}]$. The channel state estimator of Section 3.2 operates on the same single-modality sequence $z_m^{(0)}$ before pooling (or on $z_m^{(0)}$ directly for a vector modality). Instead of zero-padding an absent modality (which standard attention would still score against a query), any token whose existence is below a small threshold $\tau_{\text{miss}}$, i.e. $\alpha_m < \tau_{\text{miss}}$, is replaced once, before decoding, by a single learnable missing token $h_{\text{miss}} \in \mathbb{R}^d$ that the decoder subsequently refines from its peers. An explicitly absent input is driven to $\alpha_m \approx 0$ by the existence loss below (Section 3.4) and therefore always satisfies $\alpha_m < \tau_{\text{miss}}$.

### 3.2 Channel State Estimation

The channel state estimator (CSE) maps the single-modality encoded sequence/vector $z_m^{(0)} = \widetilde{\mathrm{Enc}}_m(y_m)$ to a triplet $q_m = (\alpha_m, \beta_m, \gamma_m)$, without access to any clean reference or to the other modalities. To lighten notation, within this subsection abbreviate the encoded sequence by $z_m = (z_{m,1}^{(0)}, \ldots, z_{m,T_m}^{(0)})^\top$ and its temporal mean by $\bar{z}_m = \frac{1}{T_m}\sum_{t=1}^{T_m} z_{m,t}^{(0)}$; for a vector modality $z_m = \bar{z}_m = h_m^{(0)}$. We now describe the three states and their signal-processing rationale.

**Existence $\alpha_m$.** An explicitly absent modality is trained to receive a small $\alpha_m$, and is treated as missing whenever $\alpha_m < \tau_{\text{miss}}$. For a present modality (channel), we combine a learned content logit with a reference-free log-energy cue in logit space,

$$\alpha_m = \sigma\Big( \underbrace{w_\alpha^\top \mathrm{ReLU}(W_\alpha \bar{z}_m + b_\alpha)}_{\text{content logit } a_\phi(\bar{z}_m)} + \underbrace{\log\big(\mathbb{E}_{t,j}[z_{m,t,j}^2] + \varepsilon\big) - \theta_E}_{\text{mean log-energy cue}} \Big) \in [0,1], \tag{3}$$

where $\mathbb{E}_{t,j}$ averages over time and feature coordinates, $\theta_E$ is a learned log-energy threshold, and $\sigma$ is the sigmoid. Eq. 3 is trained with binary cross-entropy on synthetic presence/absence patterns (Section 3.4). Using a soft probability rather than a hard flag captures *implicit* missingness: inputs that are technically present but carry no discernible signal (near-constant acoustic segments, occluded faces) have low energy and receive small $\alpha_m$, and are treated as weak channels downstream.

**Inverse-SNR proxy $\beta_m$.** Time-domain energy or variance cannot distinguish broadband corruption from semantically rich but high-variance content, so we use the shape of the spectrum/moment profile with two deterministic, parameter-free branches to define the inverse-SNR proxy $\beta_m$.

Specifically, for a sequence modality, we form the one-sided power spectrum along the time axis, averaged over the $d$ feature coordinates, $P_m(f) = \frac{1}{d}\sum_{j=1}^{d} \left|\mathrm{rFFT}_t(z_{m,\cdot,j})(f)\right|^2$, set the high band to $f \in [f_{\mathrm{hi}}, f_{\max}]$ with $f_{\mathrm{hi}} = 0.75 f_{\max}$, and use the normalized high-frequency energy ratio:

$$\beta_m = \frac{\sum_{f \geq f_{\mathrm{hi}}} P_m(f)}{\sum_f P_m(f) + \varepsilon} \in [0, 1]. \tag{4}$$

Natural speech, facial motion and text-embedding trajectories concentrate energy at low-to-mid frequencies, whereas additive broadband noise spreads energy over the whole spectrum, so $\beta_m \to 0$ for clean structured signals and $\beta_m \to 1$ for noise-dominated ones. For a vector modality (no time axis) the analogous dispersion statistic is the ratio of coordinate variance to mean-square energy,

$$\beta_m = \frac{\mathrm{Var}_j(z_{m,j})}{\mathbb{E}_j[z_{m,j}^2] + \varepsilon} \in [0, 1], \tag{5}$$

which is large for a flat, unstructured feature vector and small for a peaked, informative one. Note that $\beta_m$ (larger means noisier) is deterministic, adds no learnable parameters, and needs no training.

**Fixed channel signature and damage-type embedding $\gamma_m$.** $\alpha_m$ and $\beta_m$ cannot answer *how* a channel is corrupted, yet that matters for fusion: two channels hit by the same impairment carry correlated errors and must not corroborate each other, whereas two independently corrupted channels can still cross-correct. We first compute a fixed, 7-dimensional signature $\delta_m$ from the same spectrum/moment profile used for $\beta_m$: for a sequence it stacks the spectral centroid, spectral flatness (geometric-to-arithmetic mean power ratio), spectral kurtosis, and the relative energies of four sub-bands; for a vector it stacks mean, variance, skewness, kurtosis and range, padded to length 7. A small network then maps $\delta_m$ to a unit-norm damage-type embedding,

$$\gamma_m = \mathrm{Normalize}\big(\mathrm{MLP}_\gamma(\delta_m)\big) \in \mathbb{R}^{d_\gamma}, \qquad \|\gamma_m\|_2 = 1. \tag{6}$$

$\gamma_m$ is trained with a prototype cross-entropy: four learnable prototypes $\{c_k\}_{k=0}^{3}$ represent the synthetic damage families *clean, Gaussian, impulse, random-erasure* (the families used by the self-supervision in Section 3.4), and the embedding is pulled toward the prototype of the synthetic corruption that produced it. After training, the inner product $\gamma_m^\top \gamma_n \in [-1, 1]$ measures damage-type alignment between two channels. We show that $\gamma_m$ can recognize unseen noise in Section B.4.

### 3.3 Iterative Joint Decoding

Fusion runs $T$ shared-weight rounds, mirroring the exchange of extrinsic information between constituent decoders in a Turbo receiver (Berrou et al., 1993). Unlike a stack of distinct attention layers, the same parameters are reused at every round, so capacity grows with iterations but not with parameter count. Channel states are estimated once and held fixed across all $T$ rounds.

**Channel-modulated gate.** Before the rounds, each donor modality $n$ is assigned a reliability $r_n = \alpha_n e^{-\beta_n}$ that is high only for a present, clean channel, and each target-donor pair receives a damage-type compatibility $g_{mn} = \sigma(-\gamma_m^\top \gamma_n / \tau_\gamma)$: the minus sign means that similar damage types transfer less. The routing weight from donor modality $n$ to target $m$ ($n \neq m$) is the row-normalized product:

$$w_{n\to m} = \frac{r_n\, g_{mn}}{\sum_{k\neq m} r_k\, g_{mk} + \varepsilon} = \frac{\alpha_n e^{-\beta_n}\, \sigma(-\gamma_m^\top \gamma_n/\tau_\gamma)}{\sum_{k\neq m} \alpha_k e^{-\beta_k}\, \sigma(-\gamma_m^\top \gamma_k/\tau_\gamma) + \varepsilon}, \tag{7}$$

with self-transfers $w_{m\to m} = 0$ by masking the diagonal. The temperature $\tau_\gamma$ trades hard against soft type discrimination. Three guarantees follow directly: **(1) Gating**, $\alpha_n = 0 \Rightarrow w_{n\to m} = 0$, so an outage contributes nothing rather than a zero-padded artifact; **(2) Suppression**, the factor $e^{-\beta_n}$ shrinks a noisy donor at every round, containing error propagation during interaction rather than only at final aggregation; **(3) Decorrelation**, $g_{mn}$ decreases as two damage types align, so correlated errors cannot reinforce one another. Because routing weights are estimates, the states are detached (stop-gradient) inside the gate: the task loss cannot game the CSE by inflating $\alpha$ or shrinking $\beta$, and the CSE is trained solely through the self-supervised losses of Section 3.4.

**Channel-modulated multi-head cross-attention.** One shared block with $n_h$ heads computes queries, keys and values $Q_m = W_Q h_m$, $K_n = W_K h_n$, $V_n = W_V h_n$. For each head, a content similarity $\mathrm{softmax}_{n\neq m}(Q_m K_n^\top / \sqrt{d/n_h})$ is multiplied by the channel gate $w_{n\to m}$ and re-normalized,

so that the effective donor weight is content relevance times channel reliability,

$$\mathrm{CrossAttn}(H)_m = \mathrm{Concat}_h\Big[\sum_{n\neq m} \widetilde{a}^{(h)}_{n\to m}\, V^{(h)}_n\Big] W_O, \quad \widetilde{a}^{(h)}_{n\to m} = \frac{a^{(h)}_{n\to m}\, w_{n\to m}}{\sum_{k\neq m} a^{(h)}_{k\to m}\, w_{k\to m} + \varepsilon}, \tag{8}$$

where $a^{(h)}_{n\to m}$ is the ordinary softmax content weight (the diagonal $n = m$ is set to $-\infty$ before the softmax) and $W_O$ is the output projection. The block ends in the usual residual layer-norm and a feed-forward sublayer. Writing the resulting message to target $m$ as $p^{(t)}_m$, we have:

$$p^{(t)}_m = \mathrm{CrossAttn}\big(H^{(t-1)}\big)_m. \tag{9}$$

Unlike standard cross-modal attention (Tsai et al., 2019; Mai et al., 2025), donor messages are admitted by channel state rather than self-similarity alone (see Section A.5 for more analysis).

**Soft-information accumulation.** In a Turbo decoder, each constituent decoder passes an extrinsic-information buffer to the next, and that buffer is accumulated rather than overwritten, and therefore evidence is never lost. We instantiate the same structure with a persistent state per modality,

$$L^{(t)}_m = L^{(t-1)}_m + \eta^{(t)} p^{(t)}_m, \qquad L^{(0)}_m = \mathbf{0}, \tag{10}$$

under the geometrically decaying schedule

$$\eta^{(t)} = \eta_0\, r^t, \qquad \eta_0 > 0,\ 0 < r < 1, \tag{11}$$

which is summable: the first rounds carry the largest cross-modal corrections, while later rounds only refine, and the vanishing, summable steps guarantee that the accumulated evidence settles at a finite limit (Theorem 1).

We deliberately call $L_m$ a *soft-information state*: it plays the structural role of a Turbo log-likelihood-ratio buffer (monotone accumulation of cross-source evidence).

**Residual refinement.** The accumulated state updates the modality feature through a gated residual,

$$h^{(t)}_m = \mathrm{LayerNorm}\Big(h^{(0)}_m + W_u\, \tanh\big(W_m L^{(t)}_m\big)\Big). \tag{12}$$

The designs are load-bearing for stability and theory: the skip connection of $h^{(0)}_m$ preserves the original observation whenever cross-modal evidence is unreliable; $\tanh$ bounds the magnitude of each correction and supplies both uniform message bound and 1-Lipschitz transfer in the convergence proof (Theorem 1); and layer normalization (Ba et al., 2016) stabilizes the round-to-round scale.

**Selective reconstruction.** Only channels (modalities) flagged as low quality ($\beta_m > \tau_\beta$ *or* $\alpha_m < \tau_\alpha$) invoke a lightweight reconstructor, which avoids the common failure mode of unconditional imputation overwriting a good modality. The reconstructor averages the current peer features, maps the average through a network $R$, and blends the estimate back by exponential moving average,

$$\bar{h}^{(t)}_{-m} = \tfrac{1}{M-1} \sum_{n\neq m} h^{(t)}_n, \quad \widehat{h}^{(t)}_m = R(\bar{h}^{(t)}_{-m}), \tag{13}$$

$$h^{(t)}_m \leftarrow \begin{cases} (1-\mu)\, h^{(t)}_m + \mu\, \widehat{h}^{(t)}_m, & \beta_m > \tau_\beta \text{ or } \alpha_m < \tau_\alpha, \\ h^{(t)}_m, & \text{otherwise.} \end{cases}$$

The EMA blend retains any residual useful signal instead of replacing the modality outright. Theorem 3 shows the achievable reconstruction error shrinks precisely when present, correlated, cleaner peers are available, which is the exact situation in which Eq. 13 fires.

### 3.4 State-Weighted Fusion and Training

After $T$ rounds the final representation is a state-weighted convex combination,

$$h_{\mathrm{final}} = \sum_m \pi_m h^{(T)}_m, \qquad \pi_m = \frac{\alpha_m e^{-\beta_m}}{\sum_n \alpha_n e^{-\beta_n} + \varepsilon}, \tag{14}$$

The final prediction is: $\hat{s} = \mathrm{Predictor}(h_{\mathrm{final}})$, based on which task loss $\mathcal{L}_{\mathrm{task}}$ can be computed.

**Training objective.** The CSE needs no external clean labels: its self-supervised losses are generated on the fly from synthetic corruptions of the features. (1) **Existence**: whole modalities are randomly zeroed and the existence head is fit to the zeroing indicator by binary cross-entropy, $\mathcal{L}_{\text{exist}} = \text{BCE}(\alpha, \alpha^*)$. (2) **Damage type**: the pooled feature is corrupted with one of the four synthetic families (clean/Gaussian/impulse/erasure) and the normalized embedding is classified to its family by prototype cross-entropy, $\mathcal{L}_{\text{type}} = \text{CE}_k(\gamma^\top c_k, k^*)$. (3) **Peer reconstruction**: one modality is corrupted and the selective reconstructor must rebuild its (uncorrupted) latent feature from the remaining modalities, $\mathcal{L}_{\text{rec}} = \sum_m \|\widehat{h}_m - h_m\|_2^2$, where the target $h_m$ is the batch's own pooled representation detached from the graph. Recall that $\beta_m$ is deterministic (Eqs. 4–5) and requires no target, and these three losses are the only gradients that update the CSE. The total objective is:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_\alpha \mathcal{L}_{\text{exist}} + \lambda_\gamma \mathcal{L}_{\text{type}} + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}, \tag{15}$$

where $\mathcal{L}_{\text{task}}$ is cross-entropy (classification) or MSE (regression).

**Inference procedure.** Algorithm 1 makes the data flow explicit: modalities are first encoded, channel states and the gate are then computed once, and features are refined $T$ times.

## 4 Theoretical Analysis

We analyze one round of the decoder. Stack the modality tokens as $H = [h_1; \ldots; h_M] \in \mathbb{R}^{Md}$ and let $\Phi$ denote a full round of Eqs. 9–12. Under reasonable assumptions, the channel-modulated decoder enjoys three guarantees. We state the three guarantees informally here and give the formal assumptions, statements and proofs in Appendix A:

**(1) Stable convergence.** With a geometrically decaying accumulation schedule $\eta^{(t)} = \eta_0 r^t$ ($0 < r < 1$, see Eq. 11), the soft-information buffers in CAT-MF converge absolutely and the modality tokens converge linearly to a unique limit, $\|H^{(t)} - H^\infty\| \le K r^{t+1}$ (see Theorem 1, Eq. 20), and no step-size condition beyond $r < 1$ is needed.

**(2) Non-decreasing evidence.** Because accumulation appends each round's extrinsic messages to a slot-resolving evidence history without overwriting earlier rounds, every round adds non-negative information about the source, with a donor-level lower bound set by the best donor's existence, cross-modal correlation and cleanliness whenever the donor messages occupy separately addressable slots (Theorem 2, Eq. 26). For a Gaussian donor with noise variance $\beta_n$, the gain increases as $\beta_n$ decreases and is zero only when no present, correlated donor exists; total information is thus non-decreasing across rounds until it saturates at the joint information $I(\{h_m\}; s) \le H(s)$.

**(3) Reconstruction and noise suppression.** Whenever a present, aligned peer is available, the LMMSE reconstruction error of a weak channel decreases by an explicit amount. The decrease is larger for cleaner and more strongly aligned peers, while an absent or orthogonal peer contributes nothing, and this is exactly the condition under which the selective reconstructor fires (Theorem 3). Moreover, the gate factor $e^{-\beta_n}$ multiplies a noisy donor's propagated noise variance by a value in $[e^{-2\beta_n}, 1]$: it never amplifies noise and suppresses it by up to $e^{-2\beta_n}$ when reliable alternative donors dominate, while the single-donor case is left unchanged (Remark 1).

Also, in Section A.5 and Table 5, we summarize the advantages of CAT-MF over stacked attention.

## 5 Experiments

CAT-MF is evaluated on multimodal sentiment analysis (MSA), multimodal humor detection (MHD), multimodal sarcasm detection (MSD), Movie Genre Classification (MGC), and Handwritten Digit Classification (HDC) tasks. For complete modality evaluation, we use MOSI (Zadeh et al., 2016), MOSEI (Zadeh et al., 2018b), CH-SIMS-v2 (Liu et al., 2022), MUStARD (Castro et al., 2019), and UR-FUNNY (Hasan et al., 2019) datasets. For missing modality evaluation, we use MOSI, MOSEI, AVMNIST (Lee et al., 2017), and MM-IMDb (Vielzeuf et al., 2018) datasets. For noisy modality evaluation, we follow our baselines (Zhuang et al., 2026) to use MOSI and MOSEI.

### 5.1 Performance on Multiple Multimodal Tasks

Table 1: The results on MOSI and MOSEI. The results with † are obtained from original papers.

| Model | Venue | MOSI | | | | | MOSEI | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Acc7↑ | Acc2↑ | F1↑ | MAE↓ | Corr↑ | Acc7↑ | Acc2↑ | F1↑ | MAE↓ | Corr↑ |
| DMD (Li et al., 2023b) | CVPR 2023 | 44.9 | 84.3 | 84.3 | 0.726 | 0.788 | 52.8 | 84.6 | 84.6 | 0.538 | 0.768 |
| EMOE (Fang et al., 2025) | CVPR 2025 | 45.2 | 84.8 | 84.8 | 0.723 | 0.790 | 52.5 | 85.0 | 85.0 | 0.542 | 0.760 |
| PSA-MF (Xie et al., 2026) | AAAI 2026 | 46.5 | 86.4 | 86.2 | 0.686 | 0.807 | 55.0 | 86.3 | 86.3 | 0.521 | 0.774 |
| DHMD† (Li et al., 2026) | TPAMI 2026 | 48.4 | 86.9 | 86.9 | 0.691 | 0.800 | **55.3** | 87.1 | 87.0 | 0.539 | 0.776 |
| DecAlign (Qian et al., 2026) | ICLR 2026 | 40.9 | 86.7 | 86.3 | 0.749 | 0.790 | 52.8 | 87.0 | 86.8 | 0.534 | 0.791 |
| C-MIB (Mai et al., 2023b) | TMM 2023 | 47.7 | 87.8 | 87.8 | 0.662 | 0.835 | 52.7 | 86.9 | 86.8 | 0.542 | 0.784 |
| ITHP (Xiao et al., 2024) | ICLR 2024 | 47.7 | 88.5 | 88.5 | 0.663 | 0.856 | 52.2 | 87.1 | 87.1 | 0.550 | 0.792 |
| Diffusion Bridge (Lee et al., 2025) | CVPR 2025 | 47.3 | 86.9 | 86.8 | 0.649 | 0.839 | 53.1 | 87.1 | 87.0 | 0.531 | 0.800 |
| GSCon (Shi et al., 2025) | TIP 2025 | 45.7 | 88.1 | 88.0 | 0.696 | 0.832 | 50.8 | 87.4 | 87.4 | 0.561 | 0.750 |
| OMIB (Wu et al., 2025) | ICML 2025 | 47.6 | 88.9 | 88.8 | 0.636 | 0.844 | 53.5 | 87.3 | 87.3 | 0.530 | 0.789 |
| CAT-MF | - | **51.5** | **89.6** | **89.6** | **0.571** | **0.878** | 55.0 | **88.6** | **88.5** | **0.501** | **0.817** |

Table 3: Results on missing modality. The results marked with † are taken from (Wang et al., 2023).

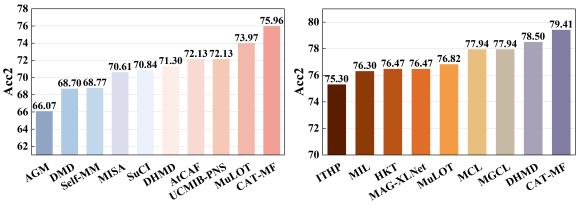
| Models | Modalities | | | | | | Avg. |
|---|---|---|---|---|---|---|---|
| | $\{l\}$ | $\{v\}$ | $\{a\}$ | $\{l,v\}$ | $\{l,a\}$ | $\{v,a\}$ | |
| **Results on MOSI** | | | | | | | |
| | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 |
| GCNet†(Lian et al., 2023) | 83.7/83.6/42.3 | 56.1/55.7/16.9 | 56.1/54.5/16.6 | 84.3/84.2/43.4 | 84.3/84.2/43.4 | 62.0/61.9/17.2 | 71.1/70.7/30.0 |
| IMDer†(Wang et al., 2023) | 84.8/84.7/44.8 | 61.3/60.8/22.2 | 62.0/62.2/22.0 | 85.5/85.4/45.3 | 85.4/85.3/45.0 | 63.6/63.4/23.8 | 73.8/73.6/33.9 |
| LNLN(Zhang et al., 2024a) | 84.9/84.7/45.1 | 52.2/58.9/18.8 | 52.2/58.9/18.0 | 84.3/84.6/44.6 | 84.9/85.2/45.1 | 52.2/58.9/18.8 | 68.5/71.9/31.7 |
| EUAR(Gao et al., 2024a) | 86.0/86.0/46.1 | 64.9/64.9/**23.6** | 63.0/62.3/**23.2** | 86.2/86.2/45.5 | 86.1/86.1/44.7 | 66.1/65.8/**24.2** | 75.4/75.2/34.5 |
| MoMKE(Xu et al., 2024) | 86.6/86.5/— | 63.4/63.3/— | 63.2/58.6/— | 87.2/87.2/— | 87.0/87.0/— | 64.0/64.7/— | 75.2/74.6/— |
| TMDC(Zhuang et al., 2026) | 87.4/87.3/— | 64.6/64.8/— | 62.4/60.2/— | 88.0/87.9/— | 87.5/87.5/— | 65.4/65.6/— | 75.9/75.5/— |
| **CAT-MF** | **89.3 /89.3 /51.0** | **65.5 / 65.5** /22.7 | **63.7 / 62.6** / 22.7 | **89.6 /89.6 /51.7** | **89.0 /89.0 /52.3** | **67.8 / 67.9** / 23.2 | **77.5/ 77.3/37.3** |
| **Results on MOSEI** | | | | | | | |
| | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 | Acc2/F1/Acc7 |
| GCNet†(Lian et al., 2023) | 83.0/83.2/51.2 | 61.9/61.6/41.7 | 60.2/60.3/41.1 | 84.3/84.4/51.1 | 84.3/84.4/51.3 | 64.1/57.2/42.0 | 73.0/71.8/46.4 |
| IMDer†(Wang et al., 2023) | 84.5/84.5/52.5 | 63.9/63.6/42.6 | 63.8/60.6/41.7 | 85.0/85.0/53.1 | 85.1/85.1/53.1 | 64.9/63.5/42.8 | 74.5/73.7/47.6 |
| LNLN(Zhang et al., 2024a) | 84.1/84.4/50.1 | 62.9/**77.2**/34.7 | 62.9/77.2/38.4 | 84.1/84.7/50.7 | 84.1/84.4/50.2 | 62.9/77.2/38.4 | 73.5/80.8/43.8 |
| EUAR(Gao et al., 2024a) | 85.3/85.2/52.9 | 66.3/65.3/42.4 | 64.5/60.7/41.6 | 86.0/86.0/53.2 | 85.1/85.1/53.7 | 66.3/65.3/42.7 | 75.6/74.6/47.8 |
| MoMKE(Xu et al., 2024) | 86.5/86.4/— | 70.1/70.2/— | 72.6/71.0/— | 86.8/86.7/— | 86.7/86.6/— | 73.3/71.8/— | 79.3/78.8/— |
| TMDC(Zhuang et al., 2026) | 86.9/86.8/— | 71.6/70.7/— | 73.6/72.2/— | 87.5/87.4/— | 87.2/87.1/— | 74.1/73.4/— | 80.1/79.6/— |
| **CAT-MF** | **87.9 /87.9 /54.3** | **77.8/77.4 /43.8** | **77.8 / 77.6/42.7** | **88.6 /88.5 /55.5** | **88.8 /88.0 /54.7** | **77.7/77.3 /43.7** | **83.1/82.8 / 49.1** |

Table 2: The results on CH-SIMS-v2.

| Model | CH-SIMS v2 | | | | | |
|---|---|---|---|---|---|---|
| | Acc5↑ | Acc3↑ | Acc2↑ | F1↑ | MAE↓ | Corr↑ |
| MISA (Hazarika et al., 2020) | 47.5 | 68.9 | 78.2 | 78.3 | 0.342 | 0.671 |
| MAG-BERT (Rahman et al., 2020) | 49.2 | 70.6 | 77.1 | 77.1 | 0.346 | 0.641 |
| Self-MM (Yu et al., 2021) | 53.5 | 72.7 | 78.7 | 78.6 | 0.315 | 0.691 |
| MMIM (Han et al., 2021) | 50.5 | 70.4 | 77.8 | 77.8 | 0.339 | 0.641 |
| AV-MC (Liu et al., 2022) | 52.1 | 73.2 | 80.6 | 80.7 | 0.301 | 0.721 |
| KuDA (Feng et al., 2024) | 53.1 | 74.3 | 80.2 | 80.1 | 0.289 | 0.741 |
| DLF (Wang et al., 2025) | 47.5 | 70.0 | 78.1 | 77.9 | 0.346 | 0.683 |
| TMDC (Zhuang et al., 2026) | 53.8 | 71.9 | 79.6 | 79.6 | 0.356 | 0.703 |
| Diffusion Bridge (Lee et al., 2025) | 52.5 | 70.7 | 78.6 | 79.0 | 0.323 | 0.677 |
| CAT-MF | **56.7** | **75.8** | **83.2** | **83.2** | **0.280** | **0.742** |

(1) **MSA task.** Table 1 reports the results on MOSI and MOSEI. On MOSI, CAT-MF ranks first on all metrics. Its Acc7 reaches 51.5, exceeding strongest competitor by 3.1 points. On MOSEI, CAT-MF is best on Acc2, F1, MAE and Corr and is a close second on Acc7, significantly outperforms OMIB. Table 2 shows the same pattern on the Chinese benchmark CH-SIMS-v2, where CAT-MF is best on every metric, with Acc2 83.2 (+2.6) and Acc5 56.7 (+2.9). **CAT-MF thus improves MSA across three datasets and two languages.** The reason is that visual and acoustic modalities often contain noise, and many baselines either compress every modality in the same way or weight modalities by feature similarity, so a weak or noisy channel is treated like a reliable one. CAT-MF instead first estimates the state of each channel (modality) and then gates how much that channel contributes, while the Turbo rounds refine all channels jointly. CAT-MF is effective even on clean data, because weak channels are turned down and degraded content is rebuilt from the reliable ones.

(2) **MHD and MSD tasks.** To test whether CAT-MF transfers across tasks, we evaluate it on MHD (UR-FUNNY dataset) and MSD (MUStARD dataset) tasks. As shown in Figure 2, CAT-MF reaches 75.96 on UR-FUNNY (+2.0 over the previous best) and 79.41 on MUStARD. Notably, we do not use the humor-centric features that several strong baselines rely on, so these gains are significant. Humor and sarcasm are largely carried by cross-modal timing: a tone of voice or a facial expression that only makes sense together with the words, and the Turbo rounds let the modalities repeatedly exchange this evidence until they agree. **These results confirm that CAT-MF generalizes beyond sentiment analysis.**




(a) Results on MHD tasks. (b) Results on MSD tasks.

Figure 2: The results on UR-FUNNY and MUStARD.

Table 4: Results on noisy modalities on MOSI and MOSEI. MuBo denotes Multimodal Boosting.

| | | Gaussian Noise | | | | | OOD Noises (Laplace Noise and Random Erasing Noise) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **NR** | **TMDC** Acc2/MAE | **C-MIB** Acc2/MAE | **MuBo** Acc2/MAE | **PDF** Acc2/MAE | **CAT-MF** Acc2/MAE | **TMDC** Acc2/MAE | **C-MIB** Acc2/MAE | **MuBo** Acc2/MAE | **PDF** Acc2/MAE | **CAT-MF** Acc2/MAE |
| MOSI | 0.1 | 87.4 / 0.748 | 87.8 / 0.670 | 86.7 / 0.678 | 87.3 / 0.766 | **89.0 / 0.651** | 87.2 / 0.769 | 87.8 / 0.666 | 87.4 / 0.639 | 87.1 / 0.841 | **88.4 / 0.649** |
| | 0.2 | 86.6 / 0.741 | 87.5 / 0.726 | 86.1 / 0.738 | 86.5 / 0.801 | **88.1 / 0.644** | 86.7 / 0.861 | 87.5 / 0.689 | 87.3 / 0.681 | 86.7 / 0.881 | **88.2 / 0.666** |
| | 0.3 | **86.9** / 0.733 | 86.4 / 0.912 | 86.4 / 0.785 | **86.9** / 0.790 | 86.7 / **0.675** | 86.8 / 0.741 | 85.1 / 1.019 | 87.1 / 0.710 | 86.1 / 0.912 | **87.5 / 0.658** |
| | 0.4 | 85.5 / 0.792 | 83.2 / 1.366 | 85.5 / 0.841 | 86.2 / 0.868 | **88.2 / 0.643** | 85.2 / 0.772 | 85.6 / 1.303 | 85.3 / 0.900 | 85.9 / 0.962 | **87.5 / 0.654** |
| | 0.5 | 85.0 / 0.912 | 84.9 / 1.660 | 86.1 / 1.172 | 85.9 / 1.112 | **87.0 / 0.672** | 84.6 / 0.867 | 83.9 / 1.900 | **86.9** / 1.114 | 85.5 / 1.023 | **86.9 / 0.682** |
| | 0.6 | 84.8 / 1.181 | 80.8 / 2.595 | 82.0 / 1.355 | **86.0** / 0.947 | 85.2 **/0.859** | 83.1 / 0.884 | 86.5 / 2.272 | 85.0 / 1.060 | 84.6 / 1.198 | **87.5 / 0.718** |
| | 0.7 | 84.0 / 1.277 | 82.1 / 3.146 | 84.4 / 1.750 | 85.5 / 1.622 | **87.2 / 0.888** | 82.0 / 0.966 | 84.0 / 3.516 | 84.7 / 1.731 | 83.6 / 1.313 | **86.7 / 0.727** |
| | **Avg** | 85.7 / 0.912 | 84.7 / 1.582 | 85.3 / 1.046 | 86.3 / 0.987 | **87.3 / 0.719** | 85.1 / 0.837 | 85.8 / 1.624 | 86.2 / 0.976 | 85.6 / 1.018 | **87.5 / 0.679** |
| MOSEI | 0.1 | 86.6 / 0.618 | 86.1 / 0.545 | 86.4 / 0.544 | 86.4 / 0.597 | **87.2 / 0.518** | 86.4 / 0.613 | 86.9 / 0.564 | 86.1 / 0.561 | 86.1 / 0.622 | **87.3 / 0.526** |
| | 0.2 | 86.2 / 0.593 | 84.5 / 0.582 | 86.6 / 0.557 | 86.0 / 0.622 | **86.9 / 0.521** | 85.5 / 0.631 | 85.1 / 0.594 | 85.2 / 0.585 | 85.7 / 0.687 | **87.3 / 0.537** |
| | 0.3 | 85.3 / 0.603 | 85.6 / 0.622 | 85.5 / 0.623 | 85.8 / 0.653 | **86.5 / 0.539** | 85.0 / 0.667 | 85.9 / 0.665 | 85.4 / 0.610 | 85.3 / 0.663 | **86.5 / 0.534** |
| | 0.4 | 84.4 / 0.596 | 84.4 / 0.703 | 85.3 / 0.682 | 85.5 / 0.688 | **86.5 / 0.539** | 84.9 / 0.682 | 84.8 / 0.767 | 84.4 / 0.714 | 84.9 / 0.801 | **86.6 / 0.533** |
| | 0.5 | 84.6 / 0.592 | 83.7 / 0.875 | 84.1 / 0.724 | 84.4 / 0.902 | **86.3 / 0.554** | 84.3 / 0.751 | 73.1 / 0.768 | 84.4 / 0.763 | 84.4 / 0.943 | **87.0 / 0.554** |
| | 0.6 | 83.8 / 0.602 | 82.4 / 1.054 | 85.4 / 0.924 | 84.7 / 0.861 | **85.7 / 0.578** | 82.7 / 0.829 | 82.0 / 0.856 | 84.1 / 0.795 | 83.5 / 0.998 | **87.0 / 0.584** |
| | 0.7 | 83.4 / 0.623 | 80.5 / 1.404 | 80.3 / 1.125 | 83.9 / 1.211 | **85.5 / 0.617** | 81.7 / 0.888 | **86.5** / 2.366 | 85.0 / 1.158 | 84.7 / 1.131 | **86.5 / 0.681** |
| | **Avg** | 84.9 / 0.604 | 83.9 / 0.826 | 84.8 / 0.740 | 85.2 / 0.791 | **86.4 / 0.552** | 84.4 / 0.723 | 83.5 / 0.940 | 84.9 / 0.741 | 84.9 / 0.835 | **86.9 / 0.564** |

### 5.2 Discussion on Missing Modalities

We next remove modalities at test time and report every possible single- and double-modality configuration. Table 3 shows that CAT-MF is best on the large majority of settings. In average Acc2, it outperforms the strongest baseline TMDC by 1.6 points on MOSI and by 3.0 points on MOSEI.

Notably, improvements are small on the language-only split, since language alone is already a strong predictor with no peer channel to exchange information with. The gains are largest where only weak modalities remain, especially on MOSEI. This follows from the channel gate, which allocates fusion weight according to the estimated channel states rather than to fixed or similarity-driven weights. When language is present, the gate already concentrates on the dominant modality. Once language is removed, it shifts weight onto the surviving weak channels and lets them corroborate each other through the iterative exchange, so a weak modality is no longer underweighted. **The fixed-missing results show that CAT-MF stays effective in modality-missing setting.** The results on MGC and HDC tasks further show that CAT-MF generalizes to new tasks and modality pairs (Section B.1).

### 5.3 Discussion on Noisy Modalities

For (1) **Gaussian noise**, we corrupt every modality with Gaussian noise, with the noise rate $NR$ ranging from 0.1 to 0.7. As Table 4 shows, CAT-MF is the best model at every noise level on both MOSI and MOSEI. **The margin widens as the noise becomes stronger.** At $NR = 0.1$ on MOSI the MAE gap is small, whereas at $NR = 0.7$ CAT-MF attains MAE 0.888 compared with 1.277 for the strongest baseline TMDC, a 30.5% reduction. Averaged over all noise levels, CAT-MF reduces MAE by 21.2% on MOSI and 8.6% on MOSEI while also leading on Acc2. For (2) **Unseen OOD noise**, we train only on Gaussian corruption and test on Laplace noise and random feature erasing that the model never sees during training. CAT-MF retains a clear advantage at every level: on average it improves Acc2 over TMDC by 2.4 on MOSI and 2.5 on MOSEI, and reduces MAE by 18.9% and 22.0%, respectively. **The gap is even larger under these unseen corruptions, showing that the robustness transfers across noise types.**

The gains stem from how most baselines handle noise: they estimate a single scalar reliability weight per modality and filter or turn down the noisy channel. This neither records the type of damage nor preserves the valid information a partially corrupted channel still contains. CAT-MF instead computes a damage-type embedding $\gamma_m$, and the channel gate can thus suppress corroboration between channels sharing the same damage while encouraging cross-checking against peers damaged in different ways. Because this factor relies on generic signal statistics rather than a Gaussian-specific formula, it also works on unseen noises. Finally, the selective reconstructor rebuilds the corrupted channel from clean peers, allowing it to recover useful content and still contribute.

## 6 Conclusion

We cast noisy and missing modalities as channel impairments and proposed CAT-MF, which estimates channel states without clean references, decodes modalities through a convergent Turbo-style iterative exchange, and gates fusion by estimated reliability and damage type. Theory guarantees linear convergence, non-decreasing information gain, and error reduction from correlated peers, and experiments confirm state-of-the-art robustness under various modality-missing and noise settings.

## AI USE STATEMENT

In this paper, AI, specifically large language models (LLMs), were used solely for language polishing to improve clarity and fluency. All scientific content, including the experimental design, results, and conclusions, is the responsibility of the authors. No LLM was involved in the ideation, analysis, or interpretation of the research, and all LLM-assisted content was carefully reviewed and verified by the authors.

## REPRODUCIBILITY STATEMENT

Code, pre-trained models, and configuration files will be released upon acceptance. All experiments use publicly available datasets and standard pre-extracted features.

## REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. In *arXiv preprint arXiv:1607.06450*, 2016.

Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *NeurIPS*, volume 32, 2019.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 59–66. IEEE, 2018.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, Feb 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2798607.

Claude Berrou, Alain Glavieux, and Punya Thitimajshima. Near shannon limit error-correcting coding and decoding: Turbo-codes. *ICC*, 2:1064–1070, 1993.

Eirina Bourtsoulatze, David Burth Kurka, and Deniz Gündüz. Deep joint source-channel coding for wireless image transmission. *IEEE Transactions on Cognitive Communications and Networking*, 5(3):567–579, 2019.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *ICML*, 2024.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *ACL*, pp. 4619–4629, 2019.

Jili Chen, Yihua Zhong, Qionghao Huang, Changqin Huang, Fan Jiang, Xiaodi Huang, and Xun Wang. Ucmib-pns: Balancing sufficiency and necessity with probabilistic causality and cross-modal uncertainty in multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 17(1):92–106, 2026. doi: 10.1109/TAFFC.2025.3606964.

Minkyoung Cho, Yulong Cao, Jiachen Sun, Qingzhao Zhang, Marco Pavone, Jeong Joon Park, Heng Yang, and Z. Morley Mao. Cocoon: Robust multi-modal perception with uncertainty-aware sensor fusion. In *International Conference on Learning Representations (ICLR)*, 2025.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep: A collaborative voice analysis repository for speech technologies. In *ICASSP*, pp. 960–964, 2014.

Reuben Dorent, Samuel Joutard, Marc Modat, Sébastien Ourselin, and Tom Vercauteren. Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 74–82. Springer, 2019.

Florian Eyben. Opensmile: the munich versatile and fast open-source audio feature extractor. In *ACM International Conference on Multimedia*, pp. 1459–1462, 2010.

Yiyang Fang, Wenke Huang, Guancheng Wan, Kehua Su, and Mang Ye. Emoe: Modality-specific enhanced dynamic emotion experts. In *CVPR*, pp. 14314–14324, 2025.

Xinyu Feng, Yuming Lin, Lihua He, You Li, Liang Chang, and Ya Zhou. Knowledge-guided dynamic modality attention fusion framework for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14755–14766, 2024.

Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *ACM MM*, pp. 9650–9659, 2024a.

Zixian Gao, Xun Jiang, Xing Xu, Fumin Shen, Yujie Li, and Heng Tao Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *CVPR*, pp. 26876–26885, 2024b.

Dimitris Gkoumas, Qiuchi Li, C. Lioma, Yijun Yu, and Da wei Song. What makes the difference? an empirical comparison of fusion strategies for multimodal language analysis. *Information Fusion*, 66:184–197, 2021.

Joachim Hagenauer, Elke Offer, and Lukas Papke. Iterative decoding of binary block and convolutional codes. *IEEE Transactions on Information Theory*, 42(2):429–445, 1996.

Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *EMNLP*, pp. 9180–9192, November 2021.

Md Kamrul Hasan, Wasifur Rahman, AmirAli Bagher Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, and Mohammed Ehsan Hoque. Ur-funny: A multimodal language dataset for understanding humor. In *EMNLP*, pp. 2046–2056, 2019.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. Humor knowledge enriched transformer for understanding multimodal humor. In *AAAI*, volume 35, pp. 12972–12980, 2021.

Devamanyu Hazarika, R. Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. *ACM MM*, pp. 1122–1131, 2020.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*, 2021.

Jiangjing Hu, Haotian Wu, Wenjing Zhang, Fengyu Wang, Wenjun Xu, Hui Gao, and Deniz Gündüz. Zero-shot semantic communication with multimodal foundation models. *arXiv preprint arXiv:2502.18200*, 2025.

Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725, 2025.

Zohar Jackson, César Souza, Jason Flaks, and Hereman Nicolas. Jakobovski/free-spoken-digit-dataset v1. 0.6. *Zenodo*, 2018.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2020.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In *CVPR*, pp. 4050–4059, 2025.

Kimin Lee, Jaehyung Kim, Song Chong, and Jinwoo Shin. Making stochastic neural networks from deterministic ones, 2017. URL `https://openreview.net/forum?id=B1akgy9xx`.

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, and Yi Zhou. Boosting multi-modal model performance with adaptive gradient modulation. In *ICCV*, pp. 22214–22224, 2023a.

Mingcheng Li, Dingkang Yang, Yunshuang Lei, Shunli Wang, Songhao Wang, Li Su, Kai Yang, Yan Wang, Ming 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, 2024.

Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24068–24077, 2025a.

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 2025b. 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/.

Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *CVPR*, pp. 6631–6640, 2023b.

Yong Li, Yuanzhi Wang, Yi Ding, Shiqing Zhang, Ke Lu, and Cuntai Guan. Decoupled hierarchical distillation for multimodal emotion recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

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.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *ACM TOMM*, 18(1):1–32, 2022.

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 *ICMI*, pp. 247–258, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

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.

Sijie Mai, Ya Sun, Ying Zeng, and Haifeng Hu. Excavating multimodal correlation for representation learning. *Information Fusion*, 91:542–555, 2023a.

Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2023b.

Sijie Mai, Ying Zeng, and Haifeng Hu. Learning from the global view: Supervised contrastive learning of multimodal representation. *Information Fusion*, 100:101920, 2023c.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14 (3):2276–2289, 2023d.

Sijie Mai, Ya Sun, Aolin Xiong, Ying Zeng, and Haifeng Hu. Multimodal boosting: Addressing noisy modalities and identifying modality contribution. *IEEE Transactions on Multimedia*, 26: 3018–3033, 2024. doi: 10.1109/TMM.2023.3306489.

Sijie Mai, Ying Zeng, Aolin Xiong, and Haifeng Hu. Injecting multimodal information into pretrained language model for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2025.

Sajad Movahedi, Vera Milovanović, Shlomo Libo Feigin, Alexander Theus, Thomas Hofmann, Valentina Boeva, T. Konstantin Rusch, and Antonio Orvieto. Fixed-point reasoners: Stable and adaptive deep looped transformers. *arXiv preprint arXiv:2606.18206*, 2026.

Jinhong Ni, Yalong Bai, Wei Zhang, Ting Yao, and Tao Mei. Deep equilibrium multimodal fusion. *arXiv preprint arXiv:2306.16645*, 2023.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. *ICML*, pp. 1310–1318, 2013.

Vittorio Pipoli, Alessia Saporita, Kevin Marchesini, Costantino Grana, Elisa Ficarra, and Federico Bolelli. Im-fuse: A mamba-based fusion block for brain tumor segmentation with incomplete modalities. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 225–235. Springer, 2025.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3930–3940, 2022.

Chengxuan Qian, Shuo Xing, Shawn Li, Yue Zhao, and Zhengzhong Tu. Decalign: Hierarchical cross-modal alignment for decoupled multimodal representation learning. In *ICLR*, 2026.

Wasifur Rahman, M. Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and E. Hoque. Integrating multimodal information in large pretrained transformers. *ACL*, 2020: 2359–2369, 2020.

Md Kaykobad Reza, Niki Nezakati, Ameya Patil, Mashhour Solh, and M. Salman Asif. Robust multimodal learning via cross-modal proxy tokens. *Transactions on Machine Learning Research*, 2025.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. In Gernot Kubin and Zdravko Kacic (eds.), *20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019*, pp. 3465–3469. ISCA, 2019. doi: 10.21437/INTERSPEECH.2019-1873. URL https://doi.org/10.21437/Interspeech.2019-1873.

QingHongYa Shi, Mang Ye, Wenke Huang, Bo Du, and Xiaofen Zong. Gradient and structure consistency in multimodal emotion recognition. *IEEE Transactions on Image Processing*, 2025.

Shankar T. Shivappa, Mohan M. Rao, and Manmohan M. Trivedi. Audiovisual information fusion in human-computer interfaces and intelligent environments: A survey. *Proceedings of the IEEE*, 98(10):1692–1715, 2010.

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 *ACL*, pp. 6558–6569, July 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pp. 0–0, 2018.

Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *AAAI*, volume 39, pp. 21180–21188, 2025.

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

Qilong Wu, Yiyang Shao, Jun Wang, and Xiaobo Sun. Learning optimal multimodal information bottleneck representations. In *International Conference on Machine Learning*, pp. 67584–67622. PMLR, 2025.

Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *Transactions on Machine Learning Research*, 2026. arXiv:2409.07825.

Xiongye Xiao, Gengshuo Liu, Gaurav Gupta, Defu Cao, Shixuan Li, Yaxing Li, Tianqing Fang, Mingxi Cheng, and Paul Bogdan. Neuro-inspired information-theoretic hierarchical perception for multimodal learning. In *ICLR*, 2024.

Heng Xie, Kang Zhu, Zhengqi Wen, Jianhua Tao, Xuefei Liu, Ruibo Fu, and Changsheng Li. Psamf: Personality-sentiment aligned multi-level fusion for multimodal sentiment analysis. In *AAAI*, 2026.

Wenxin Xu, Hexin Jiang, et al. Leveraging knowledge of modality experts for incomplete multimodal learning. In *ACM Multimedia 2024*, 2024.

Zhi Xu, Dingkang Yang, Mingcheng Li, Yuzheng Wang, Zhaoyu Chen, Jiawei Chen, Jinjie Wei, and Lihua Zhang. Debiased multimodal understanding for human language sequences. In *AAAI*, volume 39, pp. 14450–14458, 2025.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *AAAI*, volume 35, pp. 10790–10797, 2021.

Xinyue Yuan et al. M3ae: Multimodal representation learning via masked autoencoders. In *NeurIPS Workshop*, 2021.

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.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 11 2016.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *EMNLP*, pp. 1114–1125, 2017.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis Philippe Morency. Memory fusion network for multi-view sequential learning. In *AAAI*, pp. 5634–5641, 2018a.

Amir Zadeh, Paul Pu Liang, Jonathan Vanbriesen, Soujanya Poria, Edmund Tong, Erik Cambria, Minghai Chen, and Louis Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *ACL*, pp. 2236–2246, 2018b.

Zhilin Zeng, Zelin Peng, Xiaokang Yang, and Wei Shen. Missing as masking: arbitrary cross-modal feature reconstruction for incomplete multimodal brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–433. Springer, 2024.

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

Haoyu Zhang, Wenbin Wang, and Tianshu Yu. Towards robust multimodal sentiment analysis with incomplete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024a.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.*, 23(10):1499–1503, 2016. doi: 10.1109/LSP.2016.2603342. URL https://doi.org/10.1109/LSP.2016.2603342.

Mengmeng Zhang et al. Missing modality in multi-modal machine learning: A review. *arXiv preprint arXiv:2305.08772*, 2023b.

Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, and Changqing Zhang. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947*, 2024b.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. Learning multitask commonness and uniqueness for multimodal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*, 5(3):1349–1361, 2024c. doi: 10.1109/TAI.2023.3298328.

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, 2021a.

Zengqun Zhao, Qingshan Liu, and Shanmin Wang. Learning deep global multi-scale and local attention features for facial expression recognition in the wild. *IEEE Trans. Image Process.*, 30: 6544–6556, 2021b. doi: 10.1109/TIP.2021.3093397. URL https://doi.org/10.1109/TIP.2021.3093397.

Yan Zhuang, Minhao Liu, Wei Bai, Yanru Zhang, Wei Li, Jiawen Deng, and Fuji Ren. Hypermodality enhancement for multimodal sentiment analysis with missing modalities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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

**Limitations.** The states $(\alpha_m, \beta_m, \gamma_m)$ are produced from each modality in isolation and are self-supervised on four synthetic corruption families (clean, Gaussian, impulse, random erasure). The inverse-SNR proxy $\beta_m$ is a spectral/moment statistic rather than a literal SNR, so structured, semantically misleading content that happens to have a natural spectrum (stereotypical cues, demographic bias, or adversarially crafted signals) is not flagged as noise, and corruption types outside the training families may be misestimated. Moreover, the gate is computed once before the $T$ rounds and held fixed. CAT-MF therefore tracks clip-level reliability but not corruption that evolves within a sequence or across rounds. Extending the estimator to temporal, cross-modal, or semantic-level reliability is left to future work.

**Broader impact.** Multimodal systems are increasingly deployed when sensors are noisy, occluded, or intermittently unavailable. CAT-MF targets a broadly useful robustness property in this setting: an explicit, reference-free estimate of whether each modality is present and trustworthy, and a fusion rule that down-weights unreliable or damage-correlated channels instead of silently trusting them. The shared-weight turbo decoder attains iterative refinement without multiplying parameters, which lowers the training and inference cost relative to a comparably deep stacked model. We expect these properties to be beneficial in accessibility, human-computer interaction, and resource-constrained multimodal sensing.

The same technology carries risks that its reliability framing does not remove. Sentiment, humor, and sarcasm models can be applied to surveillance, profiling, or opinion monitoring, and they inherit the social and cultural biases of the pretrained language models and of datasets drawn from public YouTube, TED, and television content; a low noise score $\beta_m$ certifies statistical regularity, not fairness or factual correctness, and should not be treated as evidence that a channel is unbiased or safe to act upon. An over-confident reliability estimate could also induce misplaced trust in safety-critical decisions, and a modality that is systematically missing for a sub-population may be persistently discounted. CAT-MF is therefore not a substitute for task-specific validation, bias auditing, uncertainty calibration, and human oversight in high-stakes use. We will release code to support reproducible evaluation of multimodal robustness and encourage reporting both the reliability states and per-modality-availability results so that failure modes remain visible rather than being hidden by the fused prediction.

# A Complete Formulations and Proofs

**Notation.** Stack all modality features at round $t$ into $H^{(t)} = [h_1^{(t)}; \dots; h_M^{(t)}] \in \mathbb{R}^{Md}$, and write $\|\cdot\|$ for the Euclidean (Frobenius) norm. Let $\mathrm{Attn}_{n\to m}(H)$ denote the *content* multi-head attention message from donor $n$ to target $m$ (before channel modulation); with channel states held fixed the routing coefficients $w_{n\to m}$ of Eq. 7 are constants, and one decoding round is the map $\Phi : \mathbb{R}^{Md} \to \mathbb{R}^{Md}$ given by Eqs. 9–12:

$$
\begin{aligned}
p_m(H) &= \sum_{n \neq m} w_{n\to m}\, \mathrm{Attn}_{n\to m}(H), \\
L_m^{(t)} &= L_m^{(t-1)} + \eta^{(t)} p_m(H^{(t-1)}), \quad L_m^{(0)} = \mathbf{0}, \\
\Phi_m(H^{(t-1)}) &= \mathrm{LN}\Big(h_m^{(0)} + W_u \tanh\big(W_m L_m^{(t)}\big)\Big).
\end{aligned}
$$

This is exactly Eq. 8: multiplying the content weight $a_{n\to m}^{(h)}$ by the constant $w_{n\to m}$ inside the softmax is equivalent, at fixed states, to weighting the content message by $w_{n\to m}$ outside it. Channel states are estimated once and held fixed across the $T$ rounds, so $w_{n\to m}$ are deterministic constants in the convergence argument. The Lipschitz constant $L_{\mathrm{att}}$ below absorbs the entire shared block, including multi-head attention, output projection, residual layer-norms and the feed-forward sublayer. Under this view, the channel-modulated decoder enjoys three guarantees, which are listed as below.

## A.1 Proof of Theorem 1 (Contraction and Linear Convergence)

**Theorem 1** (Convergence and linear rate under Turbo accumulation)**.** *Suppose channel states (and hence the routing weights $w_{n\to m}$) are estimated once and held fixed across rounds. Let $C = \ell_{\mathrm{LN}} \|W_u\|_2 \|W_m\|_2$, where $\ell_{\mathrm{LN}}$ is the Lipschitz constant of layer normalization, and let the*

*soft-information schedule be geometrically decaying,*

$$\eta^{(t)} = \eta_0 r^t, \qquad \eta_0 > 0,\ 0 < r < 1. \tag{16}$$

*Since* $|\tanh(\cdot)| \le 1$ *componentwise and layer normalization maps bounded sets to bounded sets, every token* $h_m^{(t)}$ *lies in a bounded set and the shared attention block therefore produces uniformly bounded messages,*

$$B_p := \sup_{t \ge 1,\, m} \|p_m^{(t)}\| < \infty, \tag{17}$$

*where* $B_p$ *depends only on the initial tokens* $h_m^{(0)}$ *and the block weights. Then the accumulated buffers converge,*

$$L_m^{(t)} \longrightarrow L_m^{\infty} := \sum_{\tau=1}^{\infty} \eta^{(\tau)} p_m^{(\tau)}, \tag{18}$$

*and the modality tokens converge to the unique limit*

$$h_m^{\infty} = \mathrm{LN}\Big(h_m^{(0)} + W_u \tanh\big(W_m L_m^{\infty}\big)\Big). \tag{19}$$

*Moreover, for every* $t \ge 0$*,*

$$\|H^{(t)} - H^{\infty}\| \le K\, r^{t+1}, \qquad K = \frac{\sqrt{M}\, C\, B_p\, \eta_0}{1 - r}, \tag{20}$$

*i.e. convergence is linear (geometric) at rate* $r$*, and the per-round correction vanishes,* $\|H^{(t)} - H^{(t-1)}\| \le \sqrt{M}\, C\, B_p\, \eta_0\, r^t$*.*

*Proof.* We first establish uniform boundedness, then absolute convergence of the buffers, transfer it to the tokens through a Lipschitz estimate, and collect the rate.

**Step 1: uniform boundedness of tokens and messages.** The map $\tanh$ is bounded componentwise by 1, hence $\|\tanh(W_m L_m^{(t)})\|_2 \le \sqrt{d}$ for every $t$. The argument of the normalization in Eq. 12 therefore satisfies

$$\Big\| h_m^{(0)} + W_u \tanh\big(W_m L_m^{(t)}\big) \Big\|_2 \le \|h_m^{(0)}\|_2 + \|W_u\|_2 \sqrt{d},$$

which is independent of $t$. Layer normalization maps bounded sets to bounded sets, so $\sup_t \|h_m^{(t)}\| < \infty$ and the stacked state $H^{(t)}$ stays in a bounded subset of $\mathbb{R}^{Md}$. The shared block (multi-head attention, output projection, residual normalization and feed-forward sublayer) is a fixed continuous map of $H$, and on a bounded set its output is bounded; consequently the messages obey Eq. 17 for some finite $B_p$. The selective reconstruction step (Eq. 13) only forms a bounded network output of the already bounded peer tokens and blends it by an EMA, so it preserves the uniform bound.

**Step 2: absolute convergence of the buffers.** By the triangle inequality and Eq. 17,

$$\sum_{\tau=1}^{\infty} \big\|\eta^{(\tau)} p_m^{(\tau)}\big\| \le B_p \sum_{\tau=1}^{\infty} \eta_0 r^{\tau} = \frac{B_p \eta_0 r}{1 - r} < \infty.$$

The series defining $L_m^{\infty}$ therefore converges absolutely, and since $\mathbb{R}^d$ is complete, $L_m^{(t)} \to L_m^{\infty}$. The tail admits the explicit bound:

$$\|L_m^{\infty} - L_m^{(t)}\| \le \sum_{\tau=t+1}^{\infty} \eta^{(\tau)} \|p_m^{(\tau)}\| \le B_p \eta_0 \sum_{\tau=t+1}^{\infty} r^{\tau} = \frac{B_p \eta_0}{1 - r}\, r^{t+1}. \tag{21}$$

**Step 3: Lipschitz transfer to the tokens.** Define $g_m(L) := \mathrm{LN}(h_m^{(0)} + W_u \tanh(W_m L))$. The skip term $h_m^{(0)}$ cancels in a difference; using that $\tanh$ is 1-Lipschitz (its derivative $\mathrm{sech}^2$ is bounded by 1) and applying the induced-norm bounds for the linear maps,

$$\begin{aligned} \|g_m(L) - g_m(L')\| &\le \ell_{\mathrm{LN}} \big\| W_u \big(\tanh(W_m L) - \tanh(W_m L')\big) \big\| \\ &\le \ell_{\mathrm{LN}} \|W_u\|_2 \|W_m\|_2\, \|L - L'\| = C\, \|L - L'\|. \end{aligned} \tag{22}$$

Since $h_m^{(t)} = g_m(L_m^{(t)})$ and $h_m^{\infty} = g_m(L_m^{\infty})$, combining 21–22 gives

$$\|h_m^{(t)} - h_m^{\infty}\| \le \frac{CB_p\eta_0}{1-r}\, r^{t+1}.$$

Squaring, summing over $m$ and taking a square root yields Eq. 20:

$$\|H^{(t)} - H^{\infty}\| \le \frac{\sqrt{M}\, CB_p\eta_0}{1-r}\, r^{t+1} = Kr^{t+1}.$$

**Step 4: vanishing corrections and uniqueness.** The same Lipschitz estimate applied to one increment gives

$$\|h_m^{(t)} - h_m^{(t-1)}\| \le C\,\|L_m^{(t)} - L_m^{(t-1)}\| = C\,\eta^{(t)}\|p_m^{(t)}\| \le CB_p\eta_0 r^t,$$

and hence $\|H^{(t)} - H^{(t-1)}\| \le \sqrt{M}CB_p\eta_0 r^t \to 0$: the per-round correction vanishes, consistent with convergence to $H^{\infty}$. The limit is unique because it is induced by the infinite-horizon buffers, which are determined term by term from the initial tokens and the fixed channel states. This completes the proof. □

### A.2 Formulation and Proof of Theorem 2 (Per-Round Information Gain)

**Theorem 2** (Non-decreasing extrinsic evidence and per-round gain)**.** *Let* $z_{n\to m}^{(t)} := w_{n\to m}\,\mathrm{Attn}_{n\to m}(H^{(t-1)})$ *be the channel-weighted extrinsic message from donor* $n$ *to target* $m$*, and define the evidence state of* $m$ *as the slot-resolving history*

$$E_m^{(t)} = \big(h_m^{(0)},\, Z_m^{(1)}, \ldots, Z_m^{(t)}\big), \qquad Z_m^{(t)} = \big(z_{n\to m}^{(t)}\big)_{n\neq m}. \tag{23}$$

*Assume (A1)* $\eta^{(t)} > 0$ *(guaranteed by the geometric schedule of Eq. 16) and that accumulation appends evidence without overwriting earlier rounds; (A2, slot preservation) the donor messages occupy separately addressable slots, i.e.* $E_m^{(t)}$ *is in bijection with the full history of donor extrinsics rather than with a single aggregate sum; and (A3, conditional retention) some constant* $c > 0$ *satisfies, for all* $n \neq m$*,*

$$I\big(\mathrm{Attn}_{n\to m}^{(t)}; s \mid E_m^{(t-1)}\big) \ge c\, I\big(h_n^{(t-1)}; s \mid E_m^{(t-1)}\big). \tag{24}$$

*Write the conditional-retention ratio*

$$\rho_{mn}^{(t)} = \frac{I(h_n^{(t-1)}; s \mid E_m^{(t-1)})}{I(h_n^{(t-1)}; s)} \in [0,1] \quad (I(h_n^{(t-1)}; s) > 0), \tag{25}$$

*and set* $\rho_{mn}^{(t)} = 0$ *when the denominator vanishes. Then, defining the per-round gain*

$$\Delta I_m^{(t)} := I(E_m^{(t)}; s) - I(E_m^{(t-1)}; s), \tag{26}$$

*one has, for every round* $t$*,*

$$\Delta I_m^{(t)} = I\big(Z_m^{(t)}; s \mid E_m^{(t-1)}\big) \ge \max_{n\neq m} w_{n\to m}\,\rho_{mn}^{(t)}\, I(h_n^{(t-1)}; s) \ge 0. \tag{27}$$

*The equality* $\Delta I_m^{(t)} \ge 0$ *uses only the chain rule and still holds if the per-round message vector is retained but the donor slots are not. The evidence* $I(E_m^{(t)}; s)$ *is non-decreasing and bounded above by* $I(\{h_m^{(0)}\}; s) \le H(s)$*. For a Gaussian donor* $h_n = A_n s + n_n$ *with* $s \sim \mathcal{N}(\mathbf{0}, I)$*,* $n_n \sim \mathcal{N}(\mathbf{0}, \beta_n I)$ *and* $A_n A_n^{\top} = I$*,*

$$I(h_n; s) = \tfrac{1}{2}\log\det\big(I + \beta_n^{-1} A_n^{\top} A_n\big) = \tfrac{1}{2}\log(1 + 1/\beta_n), \tag{28}$$

*which is increasing in the donor's existence* $\alpha_n$*, decreasing in its noise* $\beta_n$*, and strictly positive whenever a present, conditionally correlated donor carries information beyond the current evidence state.*

*Proof.* The proof proceeds in four steps.

**Step 1: exact non-decrease of the evidence.** Because $E_m^{(t)} = (E_m^{(t-1)}, Z_m^{(t)})$ is a componentwise concatenation (the map $(E_m^{(t-1)}, Z_m^{(t)}) \mapsto E_m^{(t)}$ is a bijection with no loss of dimensions), the chain rule gives

$$\begin{aligned} I(E_m^{(t)}; s) &= I(E_m^{(t-1)}, Z_m^{(t)}; s) \\ &= I(E_m^{(t-1)}; s) + I(Z_m^{(t)}; s \mid E_m^{(t-1)}). \end{aligned} \tag{29}$$

Hence $\Delta I_m^{(t)} = I(Z_m^{(t)}; s \mid E_m^{(t-1)}) \geq 0$. If the implementation retains only the per-round aggregate $p_m^{(t)}$ instead of the donor slots, replacing $Z_m^{(t)}$ by $p_m^{(t)}$ in the same identity gives $\Delta I_m^{(t)} = I(p_m^{(t)}; s \mid E_m^{(t-1)}) \geq 0$: the non-decrease of the evidence history does not require (A2). The whole trajectory is determined term by term from $H^{(0)}$ and the fixed channel states, so $E_m^{(t)}$ is a deterministic function of $H^{(0)}$, and the data-processing inequality gives $I(E_m^{(t)}; s) \leq I(H^{(0)}; s) = I(\{h_m^{(0)}\}; s) \leq H(s)$.

**Step 2: donor-level lower bound.** Under (A2), $Z_m^{(t)}$ contains every $z_{n\to m}^{(t)}$ as a coordinate projection, so the data-processing inequality—in the correct direction, since a component is a function of the whole—gives, for each $n \neq m$,

$$I(Z_m^{(t)}; s \mid E_m^{(t-1)}) \geq I(z_{n\to m}^{(t)}; s \mid E_m^{(t-1)}). \tag{30}$$

For $w_{n\to m} > 0$, multiplication by $w_{n\to m}$ is an invertible scaling and leaves mutual information unchanged; when $w_{n\to m} = 0$ the message is identically zero and contributes nothing. Applying (A3),

$$\begin{aligned} I(z_{n\to m}^{(t)}; s \mid E_m^{(t-1)}) &= I(\mathrm{Attn}_{n\to m}^{(t)}; s \mid E_m^{(t-1)}) \\ &\geq c\, I(h_n^{(t-1)}; s \mid E_m^{(t-1)}) = c\, \rho_{mn}^{(t)} I(h_n^{(t-1)}; s). \end{aligned} \tag{31}$$

Taking the best donor yields

$$\Delta I_m^{(t)} \geq c \max_{n\neq m} w_{n\to m} \rho_{mn}^{(t)} I(h_n^{(t-1)}; s) \geq 0. \tag{32}$$

**Step 3: range of the retention ratio $\rho_{mn}^{(t)}$.** When the donor observation noise is independent of the observations already contained in $E_m^{(t-1)}$—the extrinsic conditional-independence (Gaussian) approximation standard for Turbo receivers—let $P_{\mathrm{ev}}$ be the posterior precision of $s$ induced by $E_m^{(t-1)}$ and $P_n$ the observation precision contributed by $h_n$; precisions add, so

$$\begin{aligned} I(h_n; s \mid E_m^{(t-1)}) &= \tfrac{1}{2} \log\det(I + P_{\mathrm{ev}} + P_n) - \tfrac{1}{2} \log\det(I + P_{\mathrm{ev}}) \\ &\leq \tfrac{1}{2} \log\det(I + P_n) = I(h_n; s), \end{aligned} \tag{33}$$

the inequality being the matrix form of $(1+a+b)/(1+a) \leq 1+b$, i.e. the monotonicity of $\log\det$ under the Loewner order. Thus $0 \leq \rho_{mn}^{(t)} \leq 1$. Without conditional independence $\rho$ may exceed one; in that case the donor bound remains valid with $\rho_{mn}^{(t)} I(h_n; s)$ replaced by the conditional term $I(h_n; s \mid E_m^{(t-1)})$, and no normalization is required.

**Step 4: Gaussian specialization and monotonicity.** For $h_n = A_n s + n_n$ with $A_n A_n^\top = I$,

$$I(h_n; s) = \tfrac{1}{2} \log\det(I + \beta_n^{-1} A_n^\top A_n) = \tfrac{1}{2} \log(1 + 1/\beta_n). \tag{34}$$

Write $t_n = \alpha_n g(\gamma_m, \gamma_n)$ and $S = \sum_{k\neq m,n} \alpha_k e^{-\beta_k} g(\gamma_m, \gamma_k) \geq 0$; then $w_{n\to m} = t_n e^{-\beta_n}/(t_n e^{-\beta_n} + S)$ is increasing in $\alpha_n$ and decreasing in $\beta_n$, while $\tfrac{1}{2}\log(1 + 1/\beta_n)$ is also decreasing in $\beta_n$. Hence the lower bound is (i) strictly positive when a present ($\alpha_n > 0$), conditionally correlated ($\rho_{mn}^{(t)} > 0$) donor carries information beyond the current evidence state, (ii) larger for cleaner donors, and (iii) zero when every donor is absent or redundant. □

**How this matches the implemented buffer.** Theorem 2 is stated for the evidence history, which is the object on which information is preserved by accumulation. The compressed, $d$-dimensional running buffer of Eq. 10 adds one aggregate message per round; on it, the monotonicity of Step 1 still holds for the sequence of per-round messages, whereas the donor-level bound of Step 2 and the monotonicity of the single token $I(h_m^{(t)}; s)$ are not automatic, because an aggregate sum cannot be decomposed back into its donors. To enjoy the full donor-level guarantee with the current architecture, the buffer may be split into donor-resident slots, e.g. $L_m \in \mathbb{R}^{(M-1)d}$ whose $n$-th segment accumulates $z_{n\to m}^{(t)}$; this is exactly the structure of a Turbo receiver, which keeps a separate extrinsic stream for each constituent decoder.

### A.3 Formulation and Proof of Theorem 3 (Reconstruction Error)

**Theorem 3** (Reconstruction error under correlated channels)**.** *Let $x_m = A_m s$, $s \sim \mathcal{N}(\mathbf{0}, I)$, $A_m A_m^\top = I_d$, and $y_n = A_n s + n_n$ with mutually independent $n_n \sim \mathcal{N}(\mathbf{0}, \beta_n I_d)$. Let $\mathcal{N}_m = \{n \neq m : \alpha_n > 0\}$ be the set of present peers (a hard outage, $\alpha_n = 0$, is excluded), and let*

$$\kappa_{mn} = \tfrac{1}{d}\operatorname{tr}(A_m A_n^\top) \in [-1, 1] \tag{35}$$

*be the cross-channel alignment (distinct from the retention ratio $\rho_{mn}$ in Theorem 2). In the target coordinates $s_m = P_m s$, $s_\perp = (I - P_m)s$ with $P_m = A_m^\top A_m$, partition $A_n = [B_n\ C_n]$ with $B_n = A_n P_m$, $C_n = A_n(I - P_m)$, and stack the present peers as*

$$B = \operatorname{stack}_{n\in\mathcal{N}_m} B_n, \quad C = \operatorname{stack}_{n\in\mathcal{N}_m} C_n, \quad R = \operatorname{diag}(\beta_n I_d)_{n\in\mathcal{N}_m}. \tag{36}$$

*Marginalizing the shared component $s_\perp \sim \mathcal{N}(\mathbf{0}, I)$, the peer precision about $s_m$ is*

$$P_{\text{peer}} = B^\top (R + CC^\top)^{-1} B, \tag{37}$$

*and the exact per-coordinate LMMSE error is*

$$\tfrac{1}{d}\,\text{MSE}_m = \tfrac{1}{d}\operatorname{tr}\Big[\big((1 + \alpha_m/\beta_m) I_d + P_{\text{peer}}\big)^{-1}\Big]. \tag{38}$$

*This is the irreducible Gaussian floor, and a learned reconstructor has error no smaller than it. Assume further the isotropic-alignment condition*

$$\textit{(Iso)} \qquad B_n^\top B_n \succeq \kappa_{mn}^2 I_d, \tag{39}$$

*which holds, for instance, when $A_m A_n^\top = \kappa_{mn} I_d$ (uniform alignment on the target subspace). Then the general scalar bound, valid under the shared $s_\perp$, is*

$$\tfrac{1}{d}\,\text{MSE}_m \le \left[1 + \frac{\alpha_m}{\beta_m} + \max_{n\in\mathcal{N}_m} \frac{\alpha_n \kappa_{mn}^2}{1 + \beta_n}\right]^{-1}. \tag{40}$$

*If, in addition, the peers' orthogonal components are private and conditionally independent,*

$$\textit{(C-indep)} \qquad CC^\top = \operatorname{diag}(C_n C_n^\top), \tag{41}$$

*the bound strengthens to the per-peer summation*

$$\tfrac{1}{d}\,\text{MSE}_m \le \left[1 + \frac{\alpha_m}{\beta_m} + \sum_{n\in\mathcal{N}_m} \frac{\alpha_n \kappa_{mn}^2}{1 + \beta_n}\right]^{-1}. \tag{42}$$

*The exact error decreases in the Loewner sense whenever a present, aligned peer is added; the decrease is larger for cleaner and more strongly aligned peers, and an absent or orthogonal peer contributes nothing. In the informative regime $\beta_m \ll 1$ and with no peer available, the error equals $\beta_m/(1 + \beta_m) \approx \beta_m$.*

*Proof.* The proof proceeds in four steps.

**Step 1: exact Gaussian posterior.** With $s \sim \mathcal{N}(\mathbf{0}, I_{d_s})$, $x_m = A_m s$ and $y_n = A_n s + n_n$, the stacked present observations obey $y = As + n$ with $n \sim \mathcal{N}(\mathbf{0}, R)$. Standard Gaussian conditioning gives $\Sigma_{s|y} = (I + A^\top R^{-1} A)^{-1}$ and $\mu_{s|y} = \Sigma_{s|y} A^\top R^{-1} y$; the LMMSE of $x_m = A_m s$ therefore has error

$$\text{MSE}_m = \operatorname{tr}\big(A_m \Sigma_{s|y} A_m^\top\big), \tag{43}$$

which is the smallest error attainable under the Gaussian model.

**Step 2: target coordinates and the own-channel term.** $P_m = A_m^\top A_m$ projects onto the target signal subspace $\mathcal{U}_m = \mathrm{row}(A_m)$. In the coordinates $(s_m, s_\perp)$, the target channel spans $\mathcal{U}_m$ exactly and contributes precision $\alpha_m \beta_m^{-1} I_d$ on top of the unit prior, while carrying no information about $s_\perp$. The own-only error is

$$\tfrac{1}{d}\,\mathrm{MSE}_m^{\mathrm{own}} = \big(1 + \alpha_m/\beta_m\big)^{-1} = \beta_m/(\beta_m + \alpha_m). \tag{44}$$

**Step 3: peer precision under the shared nuisance.** For a single peer, $y_n = B_n s_m + C_n s_\perp + n_n$; marginalizing $s_\perp$ gives effective noise covariance $\beta_n I_d + C_n C_n^\top$. Since $A_n A_n^\top = I_d$,

$$C_n C_n^\top = A_n (I - P_m) A_n^\top = I_d - B_n B_n^\top, \tag{45}$$

and therefore $\beta_n I_d + C_n C_n^\top = (1+\beta_n) I_d - B_n B_n^\top \preceq (1+\beta_n) I_d$, so the single-peer precision obeys

$$P_{(n)} := B_n^\top (\beta_n I_d + C_n C_n^\top)^{-1} B_n \succeq \frac{1}{1+\beta_n} B_n^\top B_n. \tag{46}$$

Under (Iso), $B_n^\top B_n \succeq \kappa_{mn}^2 I_d$, hence $P_{(n)} \succeq \kappa_{mn}^2 (1+\beta_n)^{-1} I_d$. For several peers that share one nuisance $s_\perp$, the peer precision is the joint Schur complement $P_{\mathrm{peer}} = B^\top (R + CC^\top)^{-1} B$. Adding an observation can only reduce the posterior covariance: the covariance identity

$$\mathrm{Cov}(s_m \mid y_n) = \mathbb{E}\big[\,\mathrm{Cov}(s_m \mid y_n, y_{\mathrm{rest}}) \mid y_n\big] + \mathrm{Cov}\big(\mathbb{E}[s_m \mid y_n, y_{\mathrm{rest}}] \mid y_n\big) \succeq \mathrm{Cov}(s_m \mid y_n, y_{\mathrm{rest}}) \tag{47}$$

implies, after inversion, the precision order

$$P_{\mathrm{peer}} \succeq P_{(n)} \quad \text{for every } n \in \mathcal{N}_m, \tag{48}$$

which gives the conservative bound $P_{\mathrm{peer}} \succeq \big[\max_{n\in\mathcal{N}_m} \alpha_n \kappa_{mn}^2/(1+\beta_n)\big] I_d$ and, combined with the own-channel term, Eq. 40. We stress that the per-peer terms do *not* add under the shared nuisance, because the common $s_\perp$ correlates the peers' noise; they add only under (C-indep), which makes $CC^\top$ block diagonal and yields $P_{\mathrm{peer}} = \sum_{n\in\mathcal{N}_m} P_{(n)}$, i.e. the summation bound of Eq. 42.

**Step 4: soft existence and monotonicity.** The exact expression uses hard presence $\alpha_n > 0$; the soft weights $\alpha_n$ in the scalar bounds implement the convention that existence scales the observation precision linearly, with the hard-outage endpoint $\alpha_n = 0$ exact. Every present, aligned peer increases $P_{\mathrm{peer}}$ in the Loewner order and hence cannot increase the error. The single-peer precision increases as $\beta_n$ decreases, since $(\beta_n I_d + C_n C_n^\top)^{-1}$ is Loewner-decreasing in $\beta_n$, and as the alignment $B_n$ (equivalently $|\kappa_{mn}|$) grows. An absent peer is excluded from $\mathcal{N}_m$ and an orthogonal peer has $B_n = \mathbf{0}$, $\kappa_{mn} = 0$, so either contributes nothing. □

### A.4 Formulation and Proof of Remark 1 (Noise Suppression by the Gate)

**Remark 1** (Noise suppression by the gate)**.** *Write a noisy donor $B$ as $h_B = x_B + n_B$ with $\mathbb{E}[n_B] = 0$, $\mathrm{Cov}(n_B) = \beta_B I$, and fix every other donor. Relative to the otherwise identical gate with the factor $e^{-\beta_B}$ removed, the donor's routing share is multiplied by $R = e^{-\beta_B}(t_B+S)/(t_B e^{-\beta_B}+S) \in [e^{-\beta_B}, 1]$, where $t_B = \alpha_B g(\gamma_A, \gamma_B)$ and $S$ is the total score of the other donors. Its propagated noise variance is therefore multiplied by $R^2 \in [e^{-2\beta_B}, 1]$: the gate never increases a noisy donor's share and suppresses it by* up to *$e^{-2\beta_B}$, attaining this maximum when reliable alternative donors dominate ($S \gg t_B$); when $B$ is the only available donor $R = 1$, which is unavoidable. Similarity-only attention has no $\beta_B$-dependent factor and admits the full donor noise.*

*Proof.* Consider a target $A$ and one noisy donor $B$ (the argument extends by summing over donors for $M > 2$). Decompose $h_B = x_B + n_B$ with $\mathbb{E}[n_B] = 0$, $\mathrm{Cov}(n_B) = \beta_B I$, and $n_B$ independent of the source. Because attention is linear in its value input, the noise part of $z_{B\to A} = \mathrm{Attn}_{B\to A}$ is $J_{B\to A} n_B$ for a bounded Jacobian $J$ with $\|J\|_2 \le J_{\max}$, which is held fixed across the comparison.

**Ungated / no-noise-factor baseline.** Similarity-only attention has no dependence on $\beta_B$. The fairest controlled comparison is the otherwise identical CAT-MF gate with the factor $e^{-\beta_B}$ removed (the corresponding ablation). Set $t_B := \alpha_B g(\gamma_A, \gamma_B) \ge 0$ and $S := \sum_{k\neq A,B} \alpha_k e^{-\beta_k} g(\gamma_A, \gamma_k) \ge 0$, i.e. the total score of all *other* donors. The two routing shares are

$$w_B = \frac{t_B e^{-\beta_B}}{t_B e^{-\beta_B} + S} \quad \text{(CAT-MF)}, \qquad \widetilde{w}_B = \frac{t_B}{t_B + S} \quad \text{(without } e^{-\beta_B}\text{)}.$$

**Suppression factor.** Their ratio is

$$R := \frac{w_B}{\widetilde{w}_B} = e^{-\beta_B} \frac{t_B + S}{t_B e^{-\beta_B} + S}. \tag{49}$$

Since $0 < e^{-\beta_B} \leq 1$, we have $t_B e^{-\beta_B} + S \leq t_B + S$, hence $R \leq 1$; moreover $(t_B + S)/(t_B e^{-\beta_B} + S) \leq e^{\beta_B}$ (the ratio is maximized at $S = 0$), hence $R \geq e^{-\beta_B}$. Therefore

$$e^{-\beta_B} \leq R \leq 1, \qquad e^{-2\beta_B} \leq R^2 \leq 1. \tag{50}$$

Because the noise enters linearly while its variance scales quadratically, the propagated noise variance obeys

$$\mathrm{Var}_{\mathrm{gate}} = R^2 \, \mathrm{Var}_{\text{no-factor}} \in \left[e^{-2\beta_B}, 1\right] \cdot J_{\max}^2 \beta_B, \tag{51}$$

i.e. the gate can never increase the noisy donor's share and suppresses its noise variance by up to $e^{-2\beta_B}$. The maximal suppression is attained when reliable alternative donors dominate ($S \gg t_B$); when $B$ is the only available donor ($S = 0$) one has $R = 1$, which is unavoidable (there is no alternative to fall back on) and consistent with the reconstruction result of Theorem 3. Similarity-only attention has no $\beta_B$-dependent factor and applies no such attenuation. The lower $B$'s estimated SNR (the larger $\beta_B$), the stronger the maximal suppression, which explains the widening robustness margin observed as corruption increases. □

### A.5 Comparison with stacked attention

Iterative decoding with $T$ rounds and shared weights is architecturally reminiscent of an $L{=}T$-layer cross-modal Transformer, but the two differ in four ways that are decisive under corruption. Table 5 summarizes the differences and we make each precise.

**(1) State update: accumulation vs. overwriting.** A feed-forward stack computes $h^{(\ell)} = F_\ell(h^{(\ell-1)})$ with a distinct $F_\ell$ and discards the previous representation, so evidence gathered at an early layer can be erased by a later one. CAT-MF instead keeps the running buffer $L_m^{(t)}$ of Eq. 10 by appending each round's extrinsic messages rather than overwriting them, so the corresponding evidence history retains every round's evidence. Theorem 2 shows that the information of this evidence history is therefore non-decreasing and bounded above only by the joint information $I(\{h_m\}; s) \leq H(s)$. No analogous monotonicity holds for a feed-forward stack, whose per-layer mutual information with the source can rise or fall.

**(2) Stability: summable-step convergence vs. unconstrained Jacobian.** The spectral radius of the product of layer Jacobians in a deep attention network is unconstrained. It may exceed one and amplify perturbations, the canonical training-difficulty analyzed in (Pascanu et al., 2013). In CAT-MF, the bounded $\tanh$ transfer and the skip connection keep every round of tokens in a bounded set, so the channel-modulated block produces uniformly bounded messages; under the geometrically decaying, summable schedule $\eta^{(t)} = \eta_0 r^t$, the soft-information buffers then converge absolutely and the tokens converge linearly to a limit, $\|H^{(t)} - H^{\infty}\| \leq K r^{t+1}$, with no step-size condition beyond $r < 1$ (Theorem 1). Deep equilibrium models (Bai et al., 2019) target a fixed point but neither gate their updates by input reliability nor provide a noise-propagation bound.

**(3) Robustness: state-gated vs. similarity-only mixing.** Stacked attention weights every layer from feature similarity, so a corrupted donor that stays similar to the query is admitted at every layer and its error is compounded with depth. In CAT-MF the donor weight carries $e^{-\beta_n}$ at *every* round. Remark 1 shows that, relative to the same gate without this factor, the propagated noise variance is multiplied by a factor in $[e^{-2\beta_n}, 1]$, i.e. suppressed by up to $e^{-2\beta_n}$ when reliable peers are present, and the factor $g(\gamma_m, \gamma_n)$ additionally blocks correlated damage. Depth in a standard stack multiplies exposure to corruption, while rounds in CAT-MF multiply the suppression.

**(4) Parameter budget: weight sharing vs. growth.** A stack adds a fresh parameter block per layer ($L$ times the per-layer count), whereas CAT-MF reuses one block across $T$ rounds, so its parameter count is independent of $T$ and its $O(TM^2d^2)$ compute at $T{=}L$ is matched. The comparison is therefore apples-to-apples in capacity and compute, and the gains in Table 8 for $T{=}3$ over $T{=}1$ come from the iterative, gated dynamics rather than extra parameters.

Table 5: CAT-MF decoding vs. an $L{=}T$-layer stacked cross-attention network at matched compute.

| Property | Stacked attention | CAT-MF decoding |
|---|---|---|
| State across depth | overwritten | accumulated ($L_m$, Eq. 10) |
| Information about source | not monotone | non-decreasing (Theorem. 2) |
| Convergence guarantee | none | summable-step convergence, linear rate $r$ (Theorem. 1) |
| Donor weighting | similarity only | $\alpha e^{-\beta} g$ state gate |
| Noise propagation | compounds with depth | $\times[e^{-2\beta}, 1]$ (Remark. 1) |
| Parameters vs. depth | grow $\times L$ | shared, independent of $T$ |

# B Detailed Experimental Analysis

## B.1 Results on MGC and HDC Tasks

To test generalization beyond sentiment, we transfer CAT-MF to two classification datasets with different modality pairs: MM-IMDb (text and image, multi-label genre classification) and AVMNIST (image and audio, digit classification). On MM-IMDb (Table 6), CAT-MF is best in every setting and its average F1 reaches 55.90, 1.75 points above the strongest baseline IM-Fuse. On AVMNIST (Table 7), CAT-MF is best on audio-only, full setting and average accuracy (+0.73 over SimMLM). **This confirms that CAT-MF generalizes to new tasks and new modality pairs.**

Table 6: F1-Micro (%) results on MM-IMDb. $T$ and $I$ denote Text and Image.

| Model | Setting | | | Avg. |
|---|---|---|---|---|
| | (T)(Ĩ) | (T̃)(I) | (T)(I) | |
| U-HVED (Dorent et al., 2019) | 44.26 | 39.63 | 44.04 | 42.64 |
| MaM (Zeng et al., 2024) | 54.60 | 45.00 | 56.44 | 52.01 |
| IM-Fuse (Pipoli et al., 2025) | 57.90 | 44.63 | 59.93 | 54.15 |
| SimMLM (Li et al., 2025a) | 58.74 | 42.30 | 60.35 | 53.80 |
| **CAT-MF** | **59.63** | **46.02** | **62.06** | **55.90** |

## B.2 Ablation

Table 8 removes one component at a time on MOSI. Under complete modalities every variant loses only 1.1-2.1 points, since redundant channels (modalities) already average out errors. The significance of each component grows as modalities disappear, and the pattern attributes every factor to its own failure mode:

Table 7: Accuracy (%) results on AVMNIST. $I$ and $A$ denote Image and Audio.

| Model | Setting | | | Avg. |
|---|---|---|---|---|
| | (I)(Ã) | (Ĩ)(A) | (I)(A) | |
| U-HVED (Dorent et al., 2019) | 87.67 | 92.82 | 97.70 | 92.73 |
| MaM (Zeng et al., 2024) | 89.00 | 96.74 | 99.13 | 94.96 |
| IM-Fuse (Pipoli et al., 2025) | 90.43 | 96.40 | **99.33** | 95.39 |
| SimMLM (Li et al., 2025a) | **93.40** | 94.00 | 99.07 | 95.49 |
| **CAT-MF** | 91.33 | **98.00** | **99.33** | **96.22** |

(1) **Channel state estimation and gated cross-attention (routing).** Ordinary cross-attention decides which modality to listen to purely from feature similarity, and similarity carries no information about whether a feature is trustworthy. Channel state estimation acts as a per-modality "signal-quality meter": $\alpha_m$ asks whether the channel exists at all, $\beta_m$ how clean it is, and $\gamma_m$ what kind of interference it suffers. Gated attention then turns these readings into actual volume knobs: present and clean channels are heard loudly ($\alpha_m e^{-\beta_m}$), missing channels are muted ($\alpha_m \approx 0$), noisy ones are turned down, and two channels with the same interference are prevented from confirming each other ($g(\gamma_m, \gamma_n)$). The same readings also weight the final fusion. Removing either component reverts attention to similarity-only, so a corrupted donor propagates its noise into every peer at every round and a missing slot still participates through spurious query-key matches. With all three modalities healthy this costs only 1.4/1.5 points, because averaging over three reliable channels is already safe. The loss appears once that redundancy is gone: $-2.4$ on the single-language split without the channel estimator, where un-muted missing slots might pollute the one good channel, and $-2.6$ on $\{a, v\}$ without the gate, where the noisier of the two weak channels cannot be turned down.

(2) **Damage-type factor ($g \equiv 1$).** Damage-type factor prevents two channels with the same interference from confirming each other. This variant is uniquely worst when language is missing ($-3.3$

Table 8: Ablation results (Acc2) under complete and missing modalities on MOSI.

| Variant | $\{l, a, v\}$ | $\{a, v\}$ | $\{a, l\}$ | $\{l, v\}$ | $\{l\}$ | $\{v\}$ | $\{a\}$ |
|---|---|---|---|---|---|---|---|
| Full CAT-MF | **89.6** | **67.8** | **89.0** | **89.6** | **89.3** | 65.5 | **63.7** |
| w/o channel state estimation | 88.2 | 67.1 | 88.4 | 88.5 | 86.9 | 65.4 | 63.1 |
| w/o damage-type factor ($g \equiv 1$) | 87.6 | 64.5 | 87.1 | 87.6 | 87.3 | 65.9 | 61.8 |
| w/o gated cross-modal attention | 88.1 | 65.2 | 87.8 | 87.7 | 87.1 | 66.2 | 62.7 |
| w/o iterative decoding ($T = 1$) | 87.5 | 65.6 | 86.6 | 87.7 | 85.8 | **66.5** | 61.6 |
| w/o selective reconstruction | 88.5 | 66.5 | 87.8 | 88.0 | 87.9 | 64.2 | 63.0 |

Table 9: Comparison with Stacked Cross-modal Attention on MOSI. $L$ denotes the number of attention layers.

| Operator | Parameters | Acc7 | Acc2 | F1 | MAE | Corr |
|---|---|---|---|---|---|---|
| Stack-Independent ($L$=3) | $3P$ | 48.3 | 88.1 | 88.0 | 0.583 | 0.860 |
| Stack-Shared ($L$=3) | $P$ | 48.5 | 89.0 | 89.0 | 0.584 | 0.874 |
| Stack-Independent ($L$=1) | $P$ | 48.8 | 87.6 | 87.6 | 0.615 | 0.862 |
| **CAT-MF** ($T = 3$) | $P$ | 51.5 | 89.6 | 89.6 | 0.571 | 0.878 |

on $\{a, v\}$). This is mainly because the two remaining channels (visual and acoustic) are the most likely to share the same corruption. Without $g$, they corroborate instead of cross-checking each other, which cannot be detected by a scalar reliability score because two channels can be reliably wrong together.

(3) **Iterative decoding ($T = 1$).** Iterative decoding is also very important to CAT-MF. Removing iterative decoding degrades the model performance most when all modalities are present ($-2.1$, versus at most $-2.0$ for any other factor), and the gap widens to $-3.5$ on the single-language split. This is because the iterative rounds are where cross-modal correction actually happens: a single fusion averages each modality's noisy token once, whereas Turbo rounds let every channel repeatedly absorb peer information that the peers only reveal after being refined themselves, with the soft-information buffer accumulating this evidence. One-shot fusion combines channels in one step, while decoding drives them toward a joint convergence limit. With no peers left, the same weight-tied rounds act as a recurrent denoiser, which explains the even larger single-modality loss.

(4) **Selective reconstruction.** It selectively reconstructs weak channels based on the surviving peers, which is significant as it strengthens the reliability of the channels. The cost of removing selective reconstruction tracks the number of surviving peers: $-1.1$ with all modalities and $-1.3$ to $-1.6$ whenever one is missing, matching Theorem 3, whose bound tightens with available correlated peers.

The components are thus not redundant contributors to one number: each is indispensable only under the impairment it was designed for.

### B.3 Controlled Comparison against Stacked Cross-Attention

Section A.5 discussion distinguishes CAT-MF decoding from an $L$-layer cross-modal Transformer at matched compute. To verify that the robustness advantage comes from the iterative, state-gated dynamics rather than from extra depth or parameters, we run a controlled study with four fusion operators that share identical encoders, output head, feature pipeline, optimizer, training epochs, and random seeds (only the fusion block changes).

Table 9 compares CAT-MF with stacked cross-modal attention under the same hidden width and, for the shared operators, the same parameter count:

(1) **Stacking independent layers is parameter-inefficient.** The independent $L$=3 stack uses three times the parameters ($3P$) but reaches only $88.1\%$ Acc2, a marginal $+0.5$ points over its own $L$=1 version. Its MAE improves from $0.615$ to $0.583$. Each independently learned layer transforms the representation once with nothing pulling the sequence toward a common convergence limit, so the extra depth refines the same signal only weakly and most of the added capacity is wasted.

(2) **Sharing one block across rounds turns depth into refinement.** With one third of the parameters ($P$), the shared $L{=}3$ operator reaches $89.0\%$ Acc2 and $89.0\%$ F1, outperforming the independent $3P$ stack by $+1.1$ points on Acc2 and by $+0.014$ on Corr, despite using far fewer parameters. Repeatedly applying the same learned block is equivalent to iterating a solver toward its fixed point rather than asking several distinct layers to approximate refinement separately, which converts depth from additional parameters into additional iterations of one operator.

(3) **Channel awareness closes the remaining gap.** Using identical shared rounds and the same parameter budget $P$, CAT-MF then improves over the shared stack on every reported metric: Acc7 rises by $3.0$ points, Acc2 and F1 by $0.6$ points, MAE drops to $0.571$, and Corr rises to $0.878$. The shared stack still mixes modalities from feature similarity alone, whereas CAT-MF down-weights unreliable donors at every round and selectively reconstructs degraded content, which are the two mechanisms that similarity-only rounds cannot provide.

The advantage is therefore architectural rather than a matter of size: CAT-MF with a single set of parameters ($P$) attains the best results on every metric, while the independent stack with three times as many parameters ($3P$) remains the weakest of the depth-matched operators.

### B.4 Validation of the Channel State Estimator

Because $\beta_m$ is a deterministic spectral/moment statistic while $\alpha_m$ and $\gamma_m$ are learned, we validate $\alpha_m$ and $\gamma_m$ against held-out synthetic impairments with known ground-truth state, using a bank of clean modality features. The impairments are applied to the inputs after unimodal encoders.

**(1) Does $\alpha_m$ detect whether a modality exists?** We sweep the whole-modality outage rate from $0$ to $1$ and treat the known outage as the ground-truth label. Table 10 shows that $\alpha_m$ separates present from absent channels almost perfectly (AUROC $1.00$), and the probability it assigns matches the actual fraction of present channels: both the expected calibration error and the mean-frequency tracking error are below $0.001$. In plain terms, the estimator behaves like a reliable "signal / no-signal" indicator, so the gate mutes exactly the missing channels instead of scoring their zero vectors as real tokens.

**(2) Does $\gamma_m$ identify the type of damage?** We generate one impaired version per family (clean, Gaussian, impulse, erasing), embed it with $\gamma_m$, and run a stratified 50/50 cosine nearest-neighbour test. Table 11 shows the family is read off with $99.9\%$ accuracy overall and at least $99.8\%$ for every modality. Figure 3 visualizes the same embeddings with t-SNE: the four families form four tight, well-separated clusters, while the unseen Laplace noise falls next to the Gaussian cluster, exactly as expected for two kinds of broadband noise. The embedding therefore captures similarity between damage mechanisms rather than merely recognizing the training labels.

Table 10: Existence validation of $\alpha_m$ over the outage sweep.

| Metric | Result |
|---|---|
| AUROC (present vs. absent) | 1.000 |
| Expected calibration error | 0.0001 |
| Mean-frequency tracking error | 0.0001 |

### B.5 Anti-corroboration: shared vs. independent damage

Figure 4 visualizes the donor matrix $W$ under three controlled conditions. On a clean sample, every off-diagonal entry is around $0.50$, the neutral state in which each modality trusts its peers equally. When two modalities (language and audio) are corrupted by the same damage family, the gate lowers their mutual weights to $0.39$ and redirects the released mass to the independent, uncorrupted visual modality, whose incoming and outgoing weights rise to $0.61$: two channels degraded in the same way cannot corroborate one another, so the model refuses to let them 'vouch' for each other. When the same two modalities instead carry different damage types, their mutual weights

Table 11: Damage-type identification from $\gamma_m$.

| Split | Accuracy |
|---|---|
| Overall | 99.9 |
| Language | 99.8 |
| Visual | 100.0 |
| Audio | 100.0 |

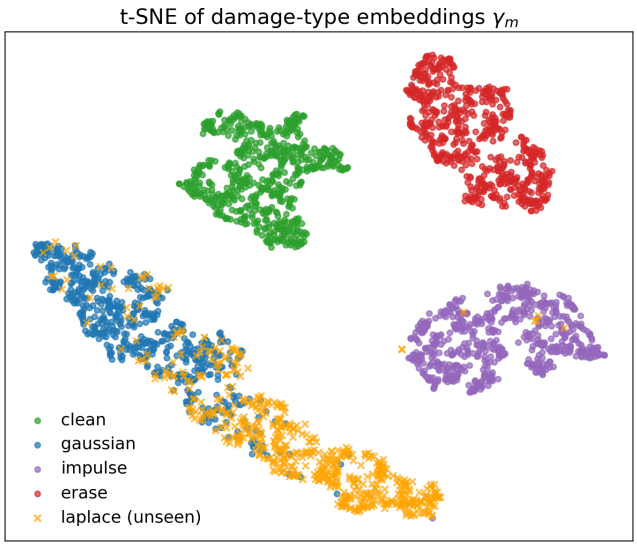



Figure 3: t-SNE of damage-type embeddings $\gamma_m$. The four warm-up families form separated clusters; Laplace noise, never seen during warm-up, lies next to Gaussian noise.

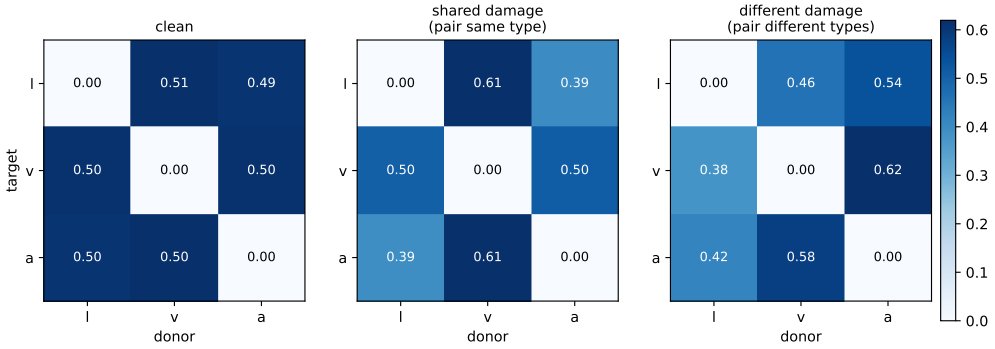



Figure 4: Donor weights under clean data, shared damage, and different damage. Shared damage suppresses transfer within the co-corrupted pair (language, audio) and redirects it to the independent visual modality. Different damage types leave the weights near their clean, balanced state.

stay close to the clean level, since differing damage signatures indicate their errors are unlikely to be shared, and cross-transfer is preserved. This routing is impossible for any scalar-reliability method, which cannot distinguish shared from independent corruption.

### B.6 Analysis of convergence guarantee

To verify the convergence guarantee of Theorem 1, we monitor the per-round update magnitude $\|H^{(t)} - H^{(t-1)}\|$ during decoding on MOSI (Figure 5). Three observations emerge. First, the first-round update grows monotonically with the severity of channel damage, from $7.0\times10^{-4}$ on clean inputs to $1.1\times10^{-3}$ at 0 dB and $1.9\times10^{-3}$ at $-5$ dB: a corrupted representation starts farther from the decoder's convergence limit and so induces a larger initial correction. Second, and crucially, the update drops by more than two orders of magnitude after a single refinement round and settles at $\sim 3\times10^{-6}$ for all three conditions, indicating that the decoder essentially reaches its convergence limit within one round and that channel damage changes the starting point but not the convergence behavior. Third, residual updates at the last round stay at the numerical floor and are negligible. These results empirically corroborate the linear-convergence and unique-limit claims of Theorem 1;

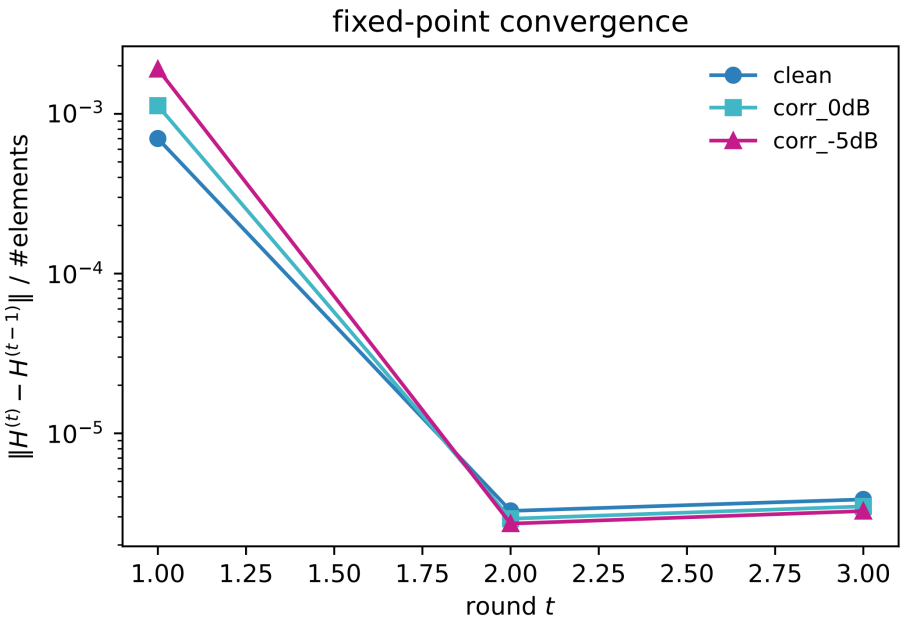
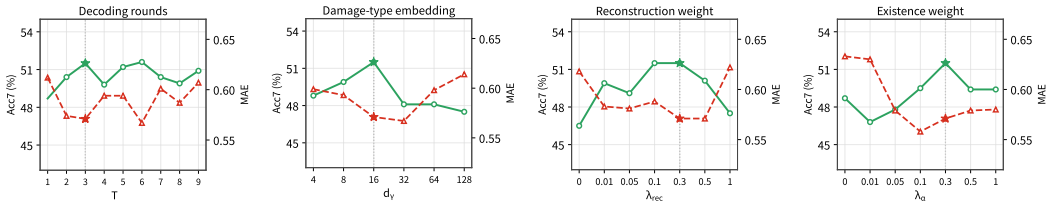



Figure 5: Convergence of the Turbo decoder on MOSI. The per-round update $\|H^{(t)} - H^{(t-1)}\|$ (averaged over elements) is shown for clean inputs and for a modality corrupted at 0 dB and $-5$ dB. Updates decay by more than two orders of magnitude within one refinement round and then remain at the numerical floor ($\sim 10^{-6}$).




(a) Decoding rounds $T$. (b) Damage-type embedding width $d_\gamma$. (c) Reconstruction weight $\lambda_{\text{rec}}$. (d) Existence weight $\lambda_\alpha$.

Figure 6: Sensitivity of CAT-MF to four key hyperparameters on MOSI. The solid line uses the left axis (Acc7, %) and the dashed line the right axis (MAE). The dotted line and stars mark the default configuration.

the observed per-round decay (two to three orders of magnitude) is faster than the theorem's conservative bound $Kr^{t+1}$, as expected for a worst-case rate.

### B.7 Hyperparameter Robustness Analysis

Figure 6 varies one hyperparameter at a time while holding all others fixed. Across all four panels Acc7 rises quickly to the default setting and is flat around it, while MAE moves in the opposite direction, and assigning extreme values to hyperparameters are harmful to the performance of the model.

(1) **Decoding rounds.** A single round is clearly the weakest, as the modalities are combined only once and never correct each other. Performance jumps at $T{=}2$ and is essentially saturated from $T{=}3$ onward, after which the tokens have reached their convergence limit and extra rounds add compute without accuracy. This empirically supports the convergence analysis and the choice $T{=}3$.

(2) **Damage-type embedding width.** Very small embeddings ($d_\gamma{=}4, 8$) cannot separate the corruption families well, while the benefit peaks at $d_\gamma{=}16$ and declines for $64, 128$: only a handful

**Algorithm 1** Channel-Aware Turbo Multimodal Fusion (inference)

**Require:** observations $\{y_m\}_{m=1}^{M}$, rounds $T$, thresholds $\tau_\alpha, \tau_\beta, \tau_{\text{miss}}$

**for** $m = 1, \dots, M$ **do**
  $z_m^{(0)} \leftarrow \widetilde{\text{Enc}}_m(y_m)$; $h_m^{(0)} \leftarrow \text{mean}_t z_{m,t}^{(0)}$ (or $z_m^{(0)}$ if a vector); $L_m^{(0)} \leftarrow \mathbf{0}$
**end for**
$\{\alpha_m, \beta_m, \gamma_m\} \leftarrow \text{CSE}(\{z_m^{(0)}\})$ ▷ Eqs. 3–6
$\{w_{n\to m}\} \leftarrow$ channel gate, Eq. 7 (computed once)
replace $h_m^{(0)}$ by $h_{\text{miss}}$ whenever $\alpha_m < \tau_{\text{miss}}$
**for** $t = 1, \dots, T$ **do**
  **for** $m = 1, \dots, M$ **do**
    $p_m^{(t)} \leftarrow$ channel-modulated MHA, Eq. 8–9
    $L_m^{(t)} \leftarrow L_m^{(t-1)} + \eta^{(t)} p_m^{(t)}$
    $h_m^{(t)} \leftarrow \text{LN}(h_m^{(0)} + W_u \tanh(W_m L_m^{(t)}))$
    **if** $\beta_m > \tau_\beta$ **or** $\alpha_m < \tau_\alpha$ **then**
      $h_m^{(t)} \leftarrow (1-\mu) h_m^{(t)} + \mu R(\frac{1}{M-1} \sum_{n \neq m} h_n^{(t)})$
    **end if**
  **end for**
**end for**
$\pi_m \leftarrow \alpha_m e^{-\beta_m} / \sum_n \alpha_n e^{-\beta_n}$; **return** $\hat{s} = \text{Predictor}(\sum_m \pi_m h_m^{(T)})$

of damage types exist, and overly wide embeddings let the damage prototypes memorize training statistics instead of generalizing. A small embedding is therefore both cheaper and better.

(3) **Reconstruction weight.** Removing reconstruction ($\lambda_{\text{rec}}$=0) gives the worst Acc7 and MAE, since the selector is never taught to repair a degraded channel. Moderate weights (0.1-0.3) work best, whereas an overly large weight (1.0) hurts again by forcing reconstruction even when the peer channels are themselves unreliable, which injects new errors.

(4) **Existence weight.** Without the existence loss ($\lambda_\alpha$=0) Acc7 is low and MAE peaks at 0.633, because missing channels are not recognized and still enter fusion through spurious matches. Accuracy peaks around 0.1-0.3 and declines at larger weights, as an over-emphasized existence signal makes the model overly conservative and mutes channels that are weak but still informative.

Overall, CAT-MF needs no extreme tuning: a small damage-type embedding, three decoding rounds, and moderate auxiliary weights already attain the optimum, and the flat region around the defaults shows that the reported results are robust to the exact hyperparameter values.

### B.8 Efficiency and Complexity Analysis

For each sample, channel estimation costs $O(Md^2)$ (dominated by the two small CSE heads; the fixed signature costs $O(MdT \log T)$ for a length-$T$ sequence) and decoding costs $O(TM^2d^2)$ time. Because weights are shared across rounds, the parameter count is independent of $T$, and the soft-information states are updated in place, so their extra memory is only $O(Md)$ rather than $O(TMd)$. With $T = 3$ the compute matches a three-layer cross-modal Transformer (Tsai et al., 2019; Mai et al., 2025) while the state memory is negligible next to the $O(M^2d^2)$ attention maps.

## C Algorithm Details

### C.1 Algorithm Pipeline

The inference pipeline of CAT-MF is shown in Algorithm 1.

### C.2 Unimodal Encoders

We first specify how each modality is mapped to the common $d$-dimensional space used by the channel-aware turbo decoder; the subscript $m \in \mathcal{M}$ indexes modalities and $d$ is the shared represen-

tation dimension (denoted $d_{per}$ in the released code). Following recent practice (Xiao et al., 2024; Hasan et al., 2021; Wang et al., 2025), the text stream is encoded with a pretrained language model (PLM) (He et al., 2021; Lan et al., 2020). Writing the raw token sequence of modality $m$ as $y_m$ and its length as $T_m$, the language front end produces

$$z_l^{(0)} = \mathrm{LN}\big(\mathrm{PLM}(y_l;\theta_l)W_{pro} + b_{pro}\big) \in \mathbb{R}^{T_l \times d}, \tag{52}$$

where $W_{pro} \in \mathbb{R}^{d_l \times d}$ and $b_{pro} \in \mathbb{R}^d$ project the PLM hidden size $d_l$ to the common dimension, and LN is layer normalization.

For the sentiment benchmarks, the acoustic and visual streams follow a temporal-convolution/Transformer front end ($m \in \{a, v\}$):

$$z_m^{(0)} = \mathrm{LN}\Big(\mathrm{Transformer}_m\big(\mathrm{Conv1D}_m(y_m;K);\theta_m\big)\Big) \in \mathbb{R}^{T_m \times d}, \tag{53}$$

where $\mathrm{Conv1D}_m$ is a temporal convolution with kernel size $K{=}3$ that first aligns the raw feature dimension, and $\mathrm{Transformer}_m$ is a shallow, modality-specific Transformer encoder (Vaswani et al., 2017). On CH-SIMS-v2, whose acoustic and visual inputs are pre-extracted feature vectors rather than sequences, the convolution/Transformer stack is replaced by a multi-layer perceptron that outputs $z_m^{(0)} \in \mathbb{R}^d$ directly.

For the humor and sarcasm benchmarks, each example contains a target punchline together with its preceding context; the two are concatenated along the time axis before encoding, and the acoustic/visual front ends apply the Transformer first and the dimension-aligning convolution second,

$$z_m^{(0)} = \mathrm{Conv1D}_m\big(\mathrm{Transformer}_m(y_m;\theta_m);K\big) \in \mathbb{R}^{T_m \times d}, \qquad m \in \{a, v\}. \tag{54}$$

For the missing-modality benchmarks (MGC and HDC tasks), the inputs are already vector features, so no temporal front end is used. On AVMNIST, the $28 \times 28$ digit image and the $20 \times 20$ MFCC spectrogram are first flattened into vectors $\tilde{y}_i \in \mathbb{R}^{784}$ and $\tilde{y}_a \in \mathbb{R}^{400}$. Each modality then passes through a single linear head followed by a projection to the common dimension,

$$z_m^{(0)} = \mathrm{LN}\big(W_{p,m}\,(W_{f,m}\tilde{y}_m) + b_{p,m}\big) \in \mathbb{R}^d, \qquad m \in \{i, a\}, \tag{55}$$

where $W_{f,i} \in \mathbb{R}^{784\times784}$ and $W_{f,a} \in \mathbb{R}^{400\times400}$ are the modality-specific feature heads, and $W_{p,m}$ projects each head output to the shared $d$-dimensional space.

On MM-IMDb, the processed release provides a pre-extracted text vector $y_l \in \mathbb{R}^{d_l}$ and poster-image vector $y_v \in \mathbb{R}^{d_v}$ (the raw dimensions $d_l, d_v$ follow the benchmark release). Each modality is mapped to the common space by a two-layer MLP,

$$z_m^{(0)} = \mathrm{LN}_2^{(m)}\Big(W_2^{(m)}\,\phi\big(\mathrm{Drop}\,\mathrm{LN}_1^{(m)}(W_1^{(m)}y_m)\big)\Big) \in \mathbb{R}^d, \qquad m \in \{l, v\}, \tag{56}$$

where $\phi$ is the GELU activation, Drop is dropout, and the intermediate width is clipped between 128 and 1024 according to the input dimension.

Because all four inputs are vector modalities, no temporal pooling is applied and the decoder tokens coincide with the encoder outputs, $h_m^{(0)} = z_m^{(0)}$, in line with Eq. 57. The channel state estimator receives these vectors directly, before any cross-modal interaction.

The decoder consumes one token per modality, obtained by temporal mean pooling of the normalized encoded sequence (a vector modality is used as-is), exactly as in Eq. 2 of the main text:

$$h_m^{(0)} = \mathrm{Enc}_m(y_m) = \begin{cases} \frac{1}{T_m}\sum_{t=1}^{T_m} z_{m,t}^{(0)} \in \mathbb{R}^d, & y_m \text{ is a sequence,} \\ z_m^{(0)} \in \mathbb{R}^d, & y_m \text{ has no time axis.} \end{cases} \tag{57}$$

Crucially, the channel state estimator of Section 3.2 receives the *pre-pooling* sequence $z_m^{(0)}$ (and the pooled $h_m^{(0)}$ for vector modalities), so the reliability, noise and damage-type estimates are formed from each modality in isolation, before any cross-modal interaction.

### C.3 Decoding and State-Estimation Modules

The unimodal tokens $H^{(0)} = [h_1^{(0)}; \ldots; h_M^{(0)}]$ then enter the three modules of Section 3: (1) The channel state estimator is a two-head network over the deterministic spectral/moment signature: a sigmoid existence head produces $\alpha_m$ and a small prototype network produces the unit-norm damage embedding $\gamma_m$, while the inverse-SNR proxy $\beta_m$ is computed without learnable parameters by Eqs. 4–5; (2) The iterative joint decoder instantiates one shared channel-modulated multi-head cross-attention block (Eq. 8), the soft-information accumulator $L_m^{(t)} = L_m^{(t-1)} + \eta^{(t)} p_m^{(t)}$ with the sigmoid warm-up of Eq. 11, and the bounded residual update of Eq. 12, in which $W_u$ is zero-initialized; the same weights are reused across all $T$ rounds; (3) The selective reconstructor is a two-layer MLP applied only to channels flagged by $\beta_m > \tau_\beta$ or $\alpha_m < \tau_\alpha$, blended by exponential moving average. The final state-weighted convex combination (Eq. 14) feeds a linear predictor. All modules are trained end-to-end in a single stage with the objective in Eq. 15; the gate and fusion detach the channel states, so the estimator is updated only through the three self-supervised terms. The design hyperparameters are listed in Table 12.

## D Experiment Settings

### D.1 Dataset Composition

We evaluate on seven benchmarks spanning sentiment, humor, sarcasm and missing-modality recognition.

**(1) MOSI** (Zadeh et al., 2016) is a widely used multimodal sentiment analysis (MSA) corpus of more than 2,000 online video clips, each annotated with a sentiment intensity in $[-3, +3]$, from strongly negative ($-3$) to strongly positive ($+3$).

**(2) MOSEI** (Zadeh et al., 2018b) is a large-scale MSA benchmark with over 22,000 clips drawn from more than 1,000 YouTube speakers and 250 topics, sampled to maximize topical and presentational diversity. Every clip carries six emotion tags and a $[-3, +3]$ sentiment score; we use the sentiment labels, which use the same scoring convention as MOSI.

**(3) CH-SIMS-v2** (Liu et al., 2022) is a Chinese MSA benchmark collected from eleven sources including interviews, talk shows and films, and designed to approximate realistic human–computer interaction. Quality filtering retains only clips with reliable audio and video. The official train/validation/test split follows a 9:2:3 ratio, comprising 2,722 (921 negative, 433 weakly negative, 232 neutral, 318 weakly positive, 818 positive), 647 and 1,034 utterances, respectively.

**(4) UR-FUNNY** (Hasan et al., 2019) targets multimodal humor detection (MHD) and is compiled from TED talks by 1,741 speakers. Each example joins a punchline utterance with its preceding context across language, audio and video; punchlines are located from audience "laughter" tags, and un-laughed target utterances form the negative class. We use version 2 (Hasan et al., 2021; Mai et al., 2023a;c), with 7,614/980/994 train/validation/test examples.

**(5) MUStARD** (Castro et al., 2019) is a multimodal sarcasm detection (MSD) corpus assembled from the TV series *Friends*, *The Big Bang Theory*, *The Golden Girls* and *Sarcasmaholics*. Its 690 utterances are labeled sarcastic or non-sarcastic and are accompanied by the preceding context utterances.

**(6) AVMNIST** (Lee et al., 2017) is a bimodal digit recognition benchmark pairing MNIST handwritten digit images (Lecun et al., 1998) with spoken digits from the Free Spoken Digit Dataset (Jackson et al., 2018). Following the missing-modality protocol of (Ma et al., 2021), the visual modality is a $28 \times 28$ grayscale image and the audio modality a $20 \times 20$ MFCC spectrogram.

**(7) MM-IMDb** (Vielzeuf et al., 2018) is a multi-label movie-genre benchmark of 25,959 poster–plot pairs over 23 genres. We use the standard 15,552/2,608/7,799 train/validation/test split; its multi-label formulation and the loose visual-textual correspondence make it the more challenging of the two missing-modality benchmarks.

### D.2 Assessment Criteria

On MOSI and MOSEI we report the five standard MSA metrics. **Acc7** rounds predictions and labels to the nearest integer in $\{-3, \ldots, 3\}$ and measures seven-class accuracy; **Acc2** and the **F1** score measure binary positive/negative discrimination and exclude neutral utterances; **MAE** is the mean absolute regression error; and **Corr** is the Pearson correlation between predicted and ground-truth scores. On CH-SIMS-v2 we additionally report Acc5 and Acc3 together with Acc2, F1, MAE and Corr, following (Liu et al., 2022; Feng et al., 2024). For MHD and MSD we report binary classification accuracy (Hasan et al., 2021; Mai et al., 2023a;c). On AVMNIST we report classification accuracy under every modality-availability configuration together with the average over configurations, and on MM-IMDb we report F1-Micro, accumulating true/false positives and false negatives over all samples and labels before computing precision, recall and F1. Both missing-modality benchmarks (AVMNIST and MM-IMDb) report results separately per availability pattern and as an average.

### D.3 Feature Extraction Strategy

**Visual.** For the MSA benchmarks we use the standard Facet descriptor set[1] of facial landmarks and action units (Mai et al., 2023d; Xiao et al., 2024); on CH-SIMS-v2 the visual features follow (Feng et al., 2024; Liu et al., 2022) and are obtained with OpenFace (Baltrusaitis et al., 2018), yielding 68 landmarks, 17 action units, head pose and orientation, and gaze. For MHD and MSD we likewise use OpenFace 2 action-unit and rigid/non-rigid shape features (Hasan et al., 2021; Mai et al., 2023a).

**Acoustic.** CH-SIMS-v2 uses 25-dimensional eGeMAPS low-level descriptors extracted with openSMILE (Eyben, 2010) at 16 kHz; the other spoken-language benchmarks use COVAREP (Degottex et al., 2014), including 12 MFCCs, pitch tracking and speech polarity.

**Language.** MOSI and MOSEI are encoded with DeBERTa (He et al., 2021; Xiao et al., 2024); CH-SIMS-v2 uses BERT (Feng et al., 2024; Liu et al., 2022); and MHD/MSD use ALBERT (Lan et al., 2020) following (Mai et al., 2023c;a).

**AVMNIST.** Images are resized to $28 \times 28$ and normalized with the standard MNIST statistics. Audio is converted to a $20 \times 20$ MFCC representation at an 8 kHz sampling rate (20 coefficients, FFT size 256, hop 80, window 256, 40 Mel filters). Image–audio pairings are fixed before training, the training portion is split into train/validation by a stratified 9:1 partition, and the test set is untouched. Complete pairs are retained at the dataset level, and missingness is imposed only through the availability indicator during training and evaluation.

**MM-IMDb.** We use the processed benchmark release, in which each movie is represented by a paired text feature vector and poster-image feature vector, both of which are normalized with training-set statistics. The official train/dev/test split is used, complete pairs are kept, and missingness is again controlled solely by the availability indicator.

### D.4 Experimental Details

CAT-MF is implemented in PyTorch 1.13.1 with CUDA 11.4 and trained on NVIDIA RTX 3090 GPUs.

**(1) Optimization and model selection.** All modules are optimized jointly in one stage with the AdamW (Loshchilov & Hutter, 2019) optimizer and a short linear learning-rate warm-up. We first delimit a practical range for each hyperparameter following (Gkoumas et al., 2021), run a 50-iteration random search on the validation set, retain the configuration with the best validation objective, and retrain five times with different random seeds, reporting the mean. Table 12 reports the per-dataset optimization settings together with the CAT-MF design hyperparameterss.

**(2) Noisy-modality protocol.** Robustness to noisy inputs is evaluated by corrupting the encoded features at test time:

$$\widetilde{h}_m^{(0)} = (1 - NR)\, h_m^{(0)} + NR\, \epsilon, \tag{58}$$

where $h_m$ is the clean (pooled) unimodal feature, $NR \in [0.1, 0.7]$ is the noise rate and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The corruption is applied feature-wise to every modality and every sample, which is the standard

[1] iMotions, 2017, `https://imotions.com/`

Table 12: Hyperparameter settings of CAT-MF.

| Hyperparameter | MOSI | MOSEI | CH-SIMS-v2 | UR-FUNNY | MUStARD | AVMNIST | MM-IMDb |
|---|---|---|---|---|---|---|---|
| Batch size | 16 | 16 | 50 | 48 | 32 | 64 | 64 |
| Learning rate | $1\times10^{-5}$ | $1\times10^{-5}$ | $1\times10^{-3}$ | $3\times10^{-6}$ | $3\times10^{-6}$ | $1\times10^{-3}$ | $1\times10^{-3}$ |
| Common dimensionality $d$ | 500 | 500 | 200 | 100 | 120 | 64 | 256 |
| Decoding rounds $T$ | 3 | 3 | 4 | 3 | 4 | 3 | 3 |
| Damage embedding dim. $d_\gamma$ | 16 | 16 | 16 | 16 | 16 | 16 | 16 |
| Type temperature $\tau_\gamma$ | 1.0 | 1.0 | 1.0 | 0.5 | 0.5 | 1.0 | 1.0 |
| $\eta_0$ | 0.5 | 0.5 | 0.25 | 0.5 | 0.8 | 0.5 | 0.5 |
| Decay ratio $r$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| Recon. existence threshold $\tau_\alpha$ | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| Recon. noise threshold $\tau_\beta$ | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |
| Missing threshold $\tau_{\text{miss}}$ | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 | 0.2 |
| EMA blend $\mu$ | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 | 0.3 |
| High-band ratio $f_{\text{hi}}/f_{\max}$ | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 | 0.75 |
| Aux. weights $(\lambda_\alpha, \lambda_\gamma, \lambda_{\text{rec}})$ | $(0.3, 0.4, 0.3)$ | $(0.3, 0.4, 0.3)$ | $(1, 1, 0.05)$ | $(1, 1, 0.1)$ | $(1, 1, 0.1)$ | $(0.2, 0.5, 0.5)$ | $(0.8, 0.8, 0.2)$ |

level at which fixed-feature multimodal benchmarks inject noise while preserving a fair and privacy-preserving protocol (Yuan et al., 2024; Mai et al., 2023b; 2024); all baselines are re-run under the identical protocol. To probe out-of-distribution corruption we further report Laplace noise and random-erasing noise (the latter zeroes a random feature subset to mimic data loss): at evaluation each test sample is corrupted, with probability $0.5$, by one of the two mechanisms.

**(3) Missing-modality protocol.** For MOSI and MOSEI datasets, fixed modality subsets are removed for every sample in a configuration. Single-modality settings drop one of $\{l, a, v\}$, and two-modality settings retain only the pair in $\{l, a\}, \{l, v\}, \{a, v\}$; e.g. $\{a, v\}$ denotes an available audio–visual pair with language absent. For a fair comparison with the missing-modality baselines (Xu et al., 2024; Zhuang et al., 2026; Lian et al., 2023), these experiments use the same feature family as our closest baselines: DeBERTa-large (He et al., 2021) gives 1024-dimensional, fine-tuned language features, wav2vec-large (Schneider et al., 2019) gives 512-dimensional acoustic features, and MTCNN face alignment (Zhang et al., 2016) followed by MA-Net (Zhao et al., 2021b) gives 1024-dimensional utterance-level visual features. The AVMNIST and MM-IMDb missing experiments use their standard availability indicators instead.

### D.5 Baselines

**Noisy-modality baselines. Multimodal Boosting** (Mai et al., 2024) is a boosting-style ensemble whose base learners cover different feature dimensions and whose contribution module estimates the importance and noise level of each learner. **C-MIB** (Mai et al., 2023b) applies the information bottleneck at both unimodal and multimodal levels to suppress redundant or noisy content. **PDF** (Cao et al., 2024) derives a predictable collaborative belief from unimodal predictions, defines single- and multi-modal confidence measures, and recalibrates the belief under uncertainty. **TMDC** (Zhuang et al., 2026) first learns clean modality-specific and shared representations from complete data and then performs inter-modal complementation for missing modalities.

**Missing-modality baselines. MMIN** (Zhao et al., 2021a) reconstructs any missing modality through cross-modal imagination. **GCNet** (Lian et al., 2023) embeds speaker and temporal graph networks for conversational missing-modality recognition and jointly trains classification and reconstruction. **IMDer** (Wang et al., 2023) uses a score-based diffusion model that turns Gaussian noise into a missing modality, conditioned on the available ones. **MoMKE** (Xu et al., 2024) trains unimodal experts first and fuses them jointly through a soft router over unimodal and combined representations. **LNLN** (Zhang et al., 2024a) combines a dominant modality correction module, a dominant-modality-based multimodal learner and a reconstructor that shields the dominant modality. **EUAR** (Gao et al., 2024a) maps each sample to a Gaussian whose variance encodes uncertainty, routes samples through a mixture of experts, and assigns them by uncertainty-aware routing.

**Additional MSA baselines. ITHP** (Xiao et al., 2024) designates one modality as the core and others as detectors that progressively filter information along an information bottleneck pathway. **Self-MM** (Yu et al., 2021) generates pseudo-labels for each modality from the multimodal label in a self-supervised manner. **DMD** (Li et al., 2023b) splits features into modality-irrelevant and -exclusive subspaces and distills them with graph units. **DLF** (Wang et al., 2025) disentangles shared from specific information under four geometric constraints and adds a language-targeted attractor. **GSCon** (Shi et al., 2025) aligns cross-modal gradient directions and preserves cross-modal

sample structure. **Diffusion Bridge** (Lee et al., 2025) bridges modalities with denoising diffusion. **EMOE** (Fang et al., 2025) combines modality experts with unimodal distillation. **DecAlign** (Qian et al., 2026) separates heterogeneous from homogeneous multimodal features in a hierarchical alignment framework. **OMIB** (Wu et al., 2025) adapts the compression strength per modality to retain task-relevant information. **AV-MC** (Liu et al., 2022) augments acoustic and visual features by modality mix-up. **MAG-BERT/MAG-ALBERT** (Rahman et al., 2020) inject acoustic and visual signals into the language model through an adaptation gate. **MISA** (Hazarika et al., 2020) decomposes each input into invariant and specific components before fusion. **MMIM** (Han et al., 2021) maximizes mutual information at several unimodal and multimodal levels. **KUDA** (Feng et al., 2024) guides fusion with emotional knowledge by dynamically selecting the dominant modality and adjusting the contributions of individual modalities. **DHMD** (Li et al., 2026) separates modality-irrelevant from modality-exclusive content and distills it through graph- and dictionary-based stages. **PSA-MF**) (Xie et al., 2026) integrates personality traits into the feature extraction process to learn personalized sentiment representations, and employs a hierarchical fusion strategy to progressively combine multimodal information, thereby enhancing sentiment analysis performance.

**Additional MHD/MSD baselines. MGCL** (Mai et al., 2023c) applies supervised contrastive learning with multi-view positives and negatives. **MCL** (Mai et al., 2023a) enforces a supervised correlation objective that preserves modality-specific structure in a joint embedding. **MIL** (Zhang et al., 2024c) jointly detects sarcasm and sentiment with target attention and a shared/specific interaction module. **HKT** (Hasan et al., 2021) injects humor-specific knowledge to resolve language ambiguity. **AGM** (Li et al., 2023a) modulates gradients according to a measured modality-competition intensity. **UCMIB-PNS** (Chen et al., 2026) pairs a variational bottleneck with a counterfactual probability-of-necessity-and-sufficiency estimator. **MuLOT** (Pramanick et al., 2022) combines self-attention with optimal-transport cross-modal matching. **MAG-XLNet** (Rahman et al., 2020) extends the multimodal adaptation gate to XLNet. **SuCI** (Xu et al., 2025) intervenes on subject-level confounders for debiased training. **AtCAF** (Huang et al., 2025) uses a front-door text-debiasing module and a counterfactual cross-modal attention.

**MGC and HDC baselines. U-HVED** (Dorent et al., 2019) jointly performs modality completion and prediction in a hetero-modal variational encoder–decoder. **MaM** (Zeng et al., 2024) treats missing modalities as masked tokens and reconstructs features under arbitrary subsets. **IM-Fuse** (Pipoli et al., 2025) is a Mamba-based incomplete-modality fusion module. **SimMLM** (Li et al., 2025a) builds a dynamic mixture of modality experts with a more-vs-fewer ranking loss that rewards adding modalities.