# Decoupled Adaptive Reconstruction for Robust Multimodal Learning under Data Corruptions

Kai Zhang, *Member, IEEE*, Mingzheng Yang, Qi Liu, *Member, IEEE*
Enhong Chen, *Fellow, IEEE*

**Abstract**—Multimodal models often fail when observations are incomplete or noisy, yet these degradations are usually addressed separately. We view both as disruptions to cross-modal redundancy and introduce **Decoupled-Adaptive Reconstruction (DAR)** for robust multimodal learning. DAR decomposes each modality into common and modality-specific representations using mutual-information objectives, restores the common component from cross-modal evidence and the specific component from intra-modal structure, and adaptively fuses restored and observed features according to reliability. Defined by information-flow roles rather than network primitives, DAR supports attention–recurrent implementations for sequential inputs and lightweight feedforward implementations for static features. We establish capacity-control and confounder-leakage bounds and derive a sufficient condition for DAR to improve test-time risk over noise-augmented learning. Experiments cover six benchmarks: MOSI, MOSEI, and SIMS under varying missing rates, and MVSA-Single, NYU Depth V2, and SUN RGB-D under Gaussian corruption. DAR achieves the strongest overall robustness in both regimes. At the highest noise level on NYU Depth V2, it improves average and worst-case accuracy over the strongest baseline by 6.17 and 8.23 percentage points, respectively. The results show that separating and restoring shared and modality-specific information provides a transferable solution to heterogeneous multimodal corruption.

**Index Terms**—Multimodal learning, multimodal fusion, data corruption, missing modalities, noisy modalities, representation disentanglement, adaptive reconstruction, mutual information.

## 1 Introduction

Multimodal learning integrates complementary information from language, vision, acoustics, depth and other sensing modalities, enabling predictions that cannot be reliably attained using only one modality. [1]–[4]. Its applications span heterogeneous input structures. In sequential tasks such as multimodal sentiment analysis (MSA), linguistic, visual, and acoustic observations must be modeled jointly over time. In static-feature tasks such as image–text classification and RGB-D scene recognition, each modality is commonly represented by a feature vector extracted from a pretrained unimodal backbone. Despite their architectural differences, most existing methods share a fundamental assumption: all modalities are available and sufficiently clean at both training and inference time.

### 1.1 A Unified View of Multimodal Corruption

The complete-and-clean assumption rarely holds in real-world deployments. Modalities may be partially unavailable because of sensor failures, packet loss, dropped frames, or missing tokens [5]–[7]. Even when a modality remains available, its observations may be degraded by image blur, ambient acoustic interference, recognition errors, or acquisition noise [8]–[11]. These degradations are usually studied as two separate problems. Missing-modality methods reconstruct unavailable features [5], [6], [12], [13], whereas noisy-modality methods estimate modality quality or down-weight unreliable predictions [8], [9], [14].

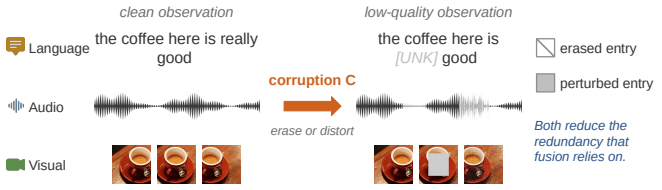



Fig. 1. A unified view of multimodal corruption. One corruption operator $C$ of (3) turns a clean multimodal observation into a low-quality one. Erasure removes part of an observation, such as a masked language token or a dropped audio span, whereas noise distorts it, as in the perturbed audio span and the corrupted video frame. The two degradations therefore differ only in how they damage the same underlying signal, and both reduce the cross-modal redundancy on which fusion relies.

We argue that missingness and noise should instead be viewed as different instances of a common data-corruption process. An erasure removes part of an observation, whereas additive or structured noise distorts it; both alter the observation associated with an underlying clean signal. More importantly, both damage the cross-modal redundancy on which multimodal fusion relies. A robust model must therefore determine which information is shared across modalities, which information is specific to each modality, and how these two types of information should be recovered from corrupted observations. This unified perspective motivates a common methodological treatment of missing and noisy inputs rather than separate solutions for each corruption type.

*Kai Zhang, Mingzheng Yang, Qi Liu, and Enhong Chen are with the State Key Laboratory of Cognitive Intelligence, University of Science and Technology of China, Hefei, Anhui 230026, China (e-mail: kkzhang08@ustc.edu.cn; yangmingzheng@mail.ustc.edu.cn; qiliuql@ustc.edu.cn; cheneh@ustc.edu.cn).*

### 1.2 Why Existing Solutions Fall Short

Direct feature-reconstruction methods provide an intuitive way to recover missing observations, but they face three limitations when applied to general multimodal corruption.

First, *entangled representations encourage shortcut reconstruction*. Methods such as TFR-Net [5] and LNLN [6] learn mappings from corrupted features to their clean counterparts without explicitly separating modality-common and modality-specific information. Consequently, their reconstruction networks may exploit incidental cross-modal correlations that are predictive in the training data but unstable under changes in corruption type or intensity. Such pseudo-shared factors can be mistaken for genuine semantic redundancy, leading to poor generalization at test time.

Second, *reconstruction and prediction objectives may interfere*. Applying a reconstruction loss directly to the encoder encourages corrupted and clean observations to have similar representations. Although this constraint can improve reconstruction accuracy, it may suppress task-discriminative information and bias the encoder toward features that are easy to reconstruct rather than features that are useful for prediction. Our diagnostic and ablation results in Sections 3.5 and 5.4 empirically demonstrate this effect.

Third, *existing solutions are often tied to a specific corruption setting or input structure*. Reconstruction models designed for missing temporal sequences do not immediately transfer to static image–text or RGB-D representations, while quality-aware fusion methods for noisy modalities generally reweight corrupted predictions without repairing the information within each modality. It therefore remains unclear whether their robustness arises from a general recovery principle or from task-specific architectural choices.

### 1.3 Decoupled-Adaptive Reconstruction

To address these limitations, we propose **Decoupled-Adaptive Reconstruction (DAR)**, a unified framework for multimodal learning under data corruptions. As illustrated in Fig. 1, DAR organizes recovery into three stages: *decouple*, *restore*, and *reconstruct*.

**Decouple.** For each modality, DAR decomposes the encoded representation into a *modal-common* component and a *modal-specific* component. The common component is encouraged to preserve information shared by corresponding observations from different modalities, whereas the specific component retains information unique to its own modality. DAR implements these constraints by maximizing cross-modal mutual information between common representations using an InfoNCE lower bound and minimizing the mutual information between common and specific representations using CLUB [15].

**Restore.** DAR restores the two components according to their distinct information sources. The modal-common component is recovered from cross-modal evidence because shared semantic information may remain observable in other modalities. The modal-specific component is recovered primarily from intra-modal structure because it cannot be inferred directly from shared information alone. A self-regression objective anchors both components to the original feature manifold and prevents arbitrary decompositions.

**Reconstruct.** An adaptive gate combines the restored components with the directly encoded observations according to their estimated reliability. This mechanism preserves trustworthy evidence while replacing information that has been severely corrupted. During training, a bootstrap branch constructed from clean features provides task-level supervision without forcing the corrupted encoder representation itself to reproduce the clean input, thereby reducing the interference between reconstruction and prediction objectives.

The same pipeline applies to both missing and noisy inputs. Missingness is modeled as an erasure of part of the observation, while noise is modeled as a distortion of that observation. In either case, DAR exposes stable cross-modal redundancy through the common representation, preserves modality-dependent evidence through the specific representation, and adaptively reconstructs the representation required by the downstream task.

DAR is also defined by the *information-flow role* of each component rather than by a fixed neural architecture. For sequential inputs, the aggregation and restoration operations are implemented using temporal self-attention, cross-attention, and recurrent estimators. For static feature vectors, the same roles are implemented using lightweight feedforward networks. Both realizations optimize the same learning objectives. This design allows us to test whether robustness originates from the decouple–restore–reconstruct principle itself rather than from a particular attention or fusion architecture (Section 4.5).

Beyond the model design, we analyze why explicit decoupling improves generalization under corruption. Under a latent-factor model containing stable common factors, modality-specific factors, and pseudo-shared confounders, we show that the mutual-information constraints control the capacity of the effective hypothesis class (Theorem 1). We further characterize two leakage paths through which a non-decoupled model can allocate representational capacity to a pseudo-shared confounder and show that DAR bounds this leakage (Theorem 2). Combining these results yields a sufficient condition under which DAR attains lower test risk than conventional noise-augmented learning (Theorem 3). Because the analysis is formulated through a general corruption operator, it applies to both erasure-based missingness and noise-based distortion.

### 1.4 Contributions

The main contributions of this work are threefold.

- First, we propose **Decoupled-Adaptive Reconstruction (DAR)**, which unifies missing and noisy multimodal observations through a decouple–restore–reconstruct framework. Defined by information-flow roles rather than specific network architectures, DAR supports both attention–recurrent and feedforward implementations with a shared training objective.
- Second, we derive a generalization analysis of DAR comprising three theorems: capacity control showing that the structural constraints never enlarge the hypothesis class (Theorem 1), a two-channel shortcut-sensitivity bound limiting how strongly the learned

representation can depend on pseudo-shared nuisance factors (Theorem 2), and a sufficient condition under which the avoided shortcut sensitivity outweighs the fitting and estimation costs, so that DAR achieves strictly lower test risk under corruption shift (Theorem 3).
- Third, we evaluate DAR on six benchmarks covering sequential and static multimodal inputs under missingness and Gaussian noise. DAR achieves strong overall robustness, including gains of 6.17 and 8.23 percentage points in average and worst-case accuracy, respectively, on NYU Depth V2 under severe noise.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the theoretical formulation and analysis. Section 4 describes DAR and its two reference implementations. Section 5 reports the experimental results. Section 6 concludes the paper.

## 2 Related Work

Research related to DAR can be grouped into three directions: robust multimodal learning with missing observations, quality-aware learning with noisy modalities, and shared–specific multimodal representation learning.

### 2.1 Robust Multimodal Learning with Missing Observations

Conventional multimodal methods typically learn joint representations by modeling intra- and inter-modal interactions. Representative approaches include tensor-based fusion [16], cross-modal Transformers [17], [18], modality translation and alignment [19], [20], and auxiliary unimodal supervision [21]. Although effective when all modalities are available, these methods generally assume complete observations and may degrade substantially when modality features are partially or entirely missing.

Recent studies explicitly improve robustness to incomplete multimodal data. This direction can be traced back to denoising autoencoders, which learn useful representations by reconstructing clean signals from corrupted inputs [22]. In the multimodal setting, SMIL [23] handles severely missing modalities through Bayesian meta-learning, and MMIN [12] imagines the representations of absent modalities with cascaded residual autoencoders. TFR-Net [5] reconstructs randomly missing elements in unaligned multimodal sequences through intra- and inter-modal attention. GCNet [7] completes missing modalities in conversational data with speaker- and temporal-aware graph neural networks. UMDF [24] employs self-distillation and dynamic feature integration to handle uncertain missing modalities, whereas LNLN [6] uses language as the dominant modality to guide the recovery of incomplete observations. Other approaches introduce adversarial training or robust representation learning to reduce sensitivity to corrupted inputs [13], adapt pretrained multimodal networks to different missing-modality configurations with a small number of learnable parameters [25], or improve the inherent robustness of multimodal Transformers through multi-task optimization and fusion-strategy search [26]. Closely related studies on incomplete multi-view learning recover unobserved views through structured joint embeddings [27] or by coupling mutual-information-based consistency learning with dual prediction [28].

Despite their effectiveness, these methods generally learn either a direct mapping from incomplete to complete features or a corruption-invariant joint representation. Consequently, modality-common information that can be recovered from other modalities and modality-specific information that must be inferred within its own modality remain entangled. DAR addresses this limitation by explicitly separating the two information sources and assigning them structurally different restoration mechanisms.

### 2.2 Quality-Aware Learning with Noisy Modalities

Another line of research adapts multimodal learning to heterogeneous modality quality. M3ER [10] uses metric learning and canonical correlation analysis to identify unreliable modalities. Noise-imitation adversarial training [13] improves representation robustness by exposing the model to generated corruptions during training. Trusted Multimodal Classification (TMC) [14] combines modality-specific evidence through Dempster–Shafer theory and was later extended with dynamic evidential fusion [29], while QMF [8] estimates modality quality to regulate multimodal optimization and fusion. DynMM [30] dynamically controls modality participation, and PDF [9] provides confidence-aware dynamic fusion for low-quality multimodal observations.

These approaches mainly improve robustness by suppressing, reweighting, or routing unreliable modality information. However, downweighting a corrupted modality may also discard task-relevant content that remains recoverable, especially when corruption affects only part of its representation. In contrast, DAR restores the latent common and specific components before fusion and then adaptively combines the observed and restored representations. Thus, it exploits the usable information within a corrupted modality instead of treating that modality as entirely unreliable.

### 2.3 Shared–Specific Multimodal Representation Learning

Multimodal observations contain both redundant information shared across modalities and complementary information unique to each modality. Explicitly separating these components has therefore become an important strategy for reducing modality heterogeneity [31]. An early factorized model [32] decomposes multimodal representations into multimodal discriminative factors and modality-specific generative factors. MISA [33] projects each modality into modality-invariant and modality-specific subspaces and aligns the invariant representations across modalities. Decoupled multimodal distillation [34] exploits shared and private representations for cross-modal knowledge transfer. ConFEDE [35] uses contrastive feature decomposition to capture similarity and discrepancy across modalities, while multimodal unified representation learning [36] adopts information-theoretic objectives to improve cross-modal generalization. ShaSpec [37] further shows that modeling shared and specific features can compensate for missing modalities in classification and segmentation tasks.

DAR builds on this shared–specific representation perspective but differs in both objective and use of the resulting factors. Specifically, DAR uses an InfoNCE lower bound [38] to preserve cross-modal information in the common representations and a CLUB upper bound [15] to reduce dependence between common and specific representations within each modality. A self-regression constraint further prevents information loss during decomposition. More importantly, the decoupled factors are not used only for fusion or regularization: common factors are restored from cross-modal redundancy, whereas specific factors are recovered from intra-modal structure. This structured restoration enables DAR to handle missingness and noise within a unified decouple–restore–reconstruct framework.

## 3 Theoretical Analysis and Synthetic Validation

This section asks *why the structural constraints imposed by DAR can improve robustness beyond training an unstructured model on corrupted observations*, focusing on the practically relevant setting in which the corruption statistics encountered at test time differ from those seen during training. Our analysis separates *stable semantic factors*, which generate the clean multimodal signal, from *environment-dependent pseudo-shared factors*, which may appear correlated across modalities during training and can therefore be exploited as shortcuts, but whose cross-modal relationships need not persist in a new corruption environment. An unconstrained corruption-augmented model can fit both types of information, whereas DAR restricts how much pseudo-shared information its common and specific channels can encode simultaneously.

The analysis proceeds in three steps: the mutual-information constraints of DAR define a restricted hypothesis class and therefore do not increase statistical complexity (Theorem 1); under an explicit two-channel coupling condition, DAR's sensitivity to pseudo-shared factors is bounded (Theorem 2); and this reduced sensitivity outweighs the induced fitting and estimation costs under an explicit sufficient condition (Theorem 3). A controlled synthetic study with known common, specific, and pseudo-shared factors then tests the observable predictions of the analysis (Section 3.5).

### 3.1 Problem Setup: Stable Semantics and Environment-Dependent Corruption

**Stable semantic factors.** Let $\mathcal{M}$ denote a finite set of modalities. For each $m \in \mathcal{M}$, the clean semantic feature is generated as

$$x_m^\star = A_c^{(m)} z_c + A_s^{(m)} \varphi\Big(z_s^{(m)}\Big) + \varepsilon_m, \tag{1}$$

where $z_c \in \mathbb{R}^{d_c}$ is a stable factor shared across modalities, $z_s^{(m)} \in \mathbb{R}^{d_s}$ is specific to modality $m$, $\varphi$ is a nonlinear transformation, and $\varepsilon_m$ denotes irreducible observation noise. The shared and specific factors together contain the semantic information that should be preserved by reconstruction.

**Pseudo-shared corruption factors.** In addition to the semantic factors, multimodal observations may contain an environment-dependent factor that appears correlated across modalities. We model this factor as

$$z_p^{(m,e)} = B_m^{(e)} z_p^{\mathrm{base}} + \rho \nu_m, \tag{2}$$

where $e \in \{\mathrm{tr}, \mathrm{te}\}$ indexes the training and test environments, $z_p^{\mathrm{base}}$ is a common nuisance source, $\nu_m$ is an independent modality-specific perturbation, and $B_m^{(e)}$ determines how the nuisance enters modality $m$ in environment $e$. When the matrices $\{B_m^{(\mathrm{tr})}\}_m$ are similar, the nuisance appears strongly correlated across modalities during training. Changes in these matrices at test time, such as sign changes, permutations, or changes in magnitude, weaken or reverse this correlation.

Unlike $z_c$, the factor $z_p^{(m,e)}$ is not part of the clean semantic target in (1). It represents recording conditions, sensor artifacts, mask patterns, or environmental disturbances that may provide an easy but unstable reconstruction shortcut.

**Unified corruption operator.** The corrupted observation in environment $e$ is written as

$$\widetilde{x}_m^{(e)} = T_m^{(e)}\left(x_m^\star, \xi_m^{(e)}, z_p^{(m,e)}\right), \tag{3}$$

where $\xi_m^{(e)}$ contains the remaining stochastic corruption variables. For local analysis, the operator can be represented as

$$\widetilde{x}_m^{(e)} = C_m\Big(\xi_m^{(e)}\Big)\, x_m^\star + A_p^{(m)} z_p^{(m,e)} + \omega_m^{(e)}. \tag{4}$$

For missing observations, $C_m(\xi_m) = \mathrm{diag}(\xi_m)$ with a binary mask $\xi_m$, whereas for additive corruption $C_m(\xi_m) = I$ and $\omega_m^{(e)}$ contains the modality-specific noise.

Although $z_p^{(m,e)}$ enters through a single mixing matrix $A_p^{(m)}$, its effect generally overlaps both the common and specific semantic subspaces. Let $\Pi_c^{(m)}$ and $\Pi_s^{(m)}$ denote the corresponding orthogonal projections. The two observable nuisance channels are

$$\begin{aligned} A_{p,c}^{(m)} &= \Pi_c^{(m)} A_p^{(m)}, \\ A_{p,s}^{(m)} &= \Pi_s^{(m)} A_p^{(m)}. \end{aligned} \tag{5}$$

The first channel can be mistaken for cross-modal semantic redundancy, whereas the second can be absorbed as modality-specific information. This two-channel overlap motivates the common–specific constraints used by DAR.

**Reconstruction risk.** Let

$$f : (\widetilde{x}_m)_{m\in\mathcal{M}} \mapsto (\widehat{x}_m)_{m\in\mathcal{M}}$$

be a reconstruction map. Its population risk in environment $e$ is

$$R_e(f) = \mathbb{E}_{P_e} \sum_{m\in\mathcal{M}} \Big\| f_m(\widetilde{x}^{(e)}) - x_m^\star \Big\|_2^2. \tag{6}$$

The corresponding empirical training risk is

$$\widehat{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \sum_{m\in\mathcal{M}} \Big\| f_m(\widetilde{x}^{(i,\mathrm{tr})}) - x_m^{\star(i)} \Big\|_2^2. \tag{7}$$

The theory below analyzes reconstruction risk. If the downstream predictor and its loss are Lipschitz, the resulting reconstruction-risk bounds transfer directly to downstream prediction risk up to the corresponding Lipschitz constants.

### 3.2 Structured and Unstructured Hypothesis Classes

We compare DAR with an unstructured model trained on samples generated by the same corruption process used for DAR during training.

**Definition 1** (Corruption-augmented hypothesis class)**.** *Let $\mathcal{H}_{\mathrm{aug}}$ be the class of admissible reconstruction maps with bounded parameters and Lipschitz constant:*

$$\mathcal{H}_{\mathrm{aug}} = \{f_\theta : \|\theta\| \le B,\ f_\theta\ L_f\text{-Lipschitz}\}. \tag{8}$$

*Its empirical minimizer is*

$$\widehat{f}_{\mathrm{aug}} \in \arg\min_{f\in\mathcal{H}_{\mathrm{aug}}} \widehat{R}_n(f). \tag{9}$$

The class $\mathcal{H}_{\mathrm{aug}}$ makes no distinction between common and specific information. It is sufficiently expressive to include structured solutions, but empirical risk minimization does not explicitly discourage solutions that rely on pseudo-shared factors.

**Definition 2** (DAR hypothesis class)**.** *The DAR class is the subset of $\mathcal{H}_{\mathrm{aug}}$ admitting modality-wise common and specific encoders:*

$$\begin{aligned}\mathcal{H}_{\mathrm{DAR}} = \Big\{ f \in \mathcal{H}_{\mathrm{aug}} :\ & f(\widetilde{x}) = \mathcal{D}\big(\{E_c^{(m)}(\widetilde{x}_m), E_s^{(m)}(\widetilde{x}_m)\}_m\big),\\ & I_{\mathrm{tr}}\big(E_c^{(m)}; E_s^{(m)}\big) \le \varepsilon_d \quad \forall m,\\ & I_{\mathrm{tr}}\big(E_c^{(m)}; E_c^{(m')}\big) \ge \varepsilon_s \quad \forall m \ne m' \Big\}.\end{aligned} \tag{10}$$

*Its empirical minimizer is denoted by*

$$\widehat{f}_{\mathrm{DAR}} \in \arg\min_{f\in\mathcal{H}_{\mathrm{DAR}}} \widehat{R}_n(f).$$

The first constraint reduces dependence between the common and specific channels within each modality, whereas the second preserves cross-modal semantic consistency. Throughout, $I_{\mathrm{tr}}(E_c^{(m)}; E_s^{(m)})$ abbreviates the mutual information between the *outputs* $E_c^{(m)}(\widetilde{x}_m)$ and $E_s^{(m)}(\widetilde{x}_m)$ under the training corruption distribution, and similarly for $I_{\mathrm{tr}}(E_c^{(m)}; E_c^{(m')})$. By construction,

$$\mathcal{H}_{\mathrm{DAR}} \subseteq \mathcal{H}_{\mathrm{aug}}. \tag{11}$$

Consequently, the unstructured class can achieve no larger empirical training error:

$$\widehat{\Delta}_n := \widehat{R}_n(\widehat{f}_{\mathrm{DAR}}) - \widehat{R}_n(\widehat{f}_{\mathrm{aug}}) \ge 0. \tag{12}$$

The quantity $\widehat{\Delta}_n$ is the empirical fitting price paid for the structural constraints. The purpose of the analysis is to determine when this price is offset by improved generalization and lower sensitivity to corruption shift.

### 3.3 Assumptions

**Assumption 1** (Regularity)**.** *The latent variables are sub-Gaussian, all mixing matrices have bounded operator norm, and the outputs of the encoders and decoder lie in bounded sets. The reconstruction loss is bounded by $M_\ell$ and is $L_\ell$-Lipschitz with respect to the reconstructed output.*

**Assumption 2** (Stable semantic identifiability)**.** *The shared factor $z_c$ and the modality-specific factors $\{z_s^{(m)}\}_m$ are invariant across corruption environments. Moreover, $z_c$ is the only factor that is simultaneously shared across modalities and stable across the training corruption environments. The residual error in identifying this stable common subspace is bounded by $\eta_s \ge 0$.*

To formulate the two-channel condition, define the average nuisance sensitivities of the two encoders as

$$\begin{aligned}\lambda_{p,c}^{(m)2} &= \mathbb{E}\left\|\partial_{z_p} E_c^{(m)}(\widetilde{x}_m)\right\|_F^2,\\ \lambda_{p,s}^{(m)2} &= \mathbb{E}\left\|\partial_{z_p} E_s^{(m)}(\widetilde{x}_m)\right\|_F^2.\end{aligned} \tag{13}$$

**Assumption 3** (Balanced two-channel coupling)**.** *For the pseudo-shared factors considered in the analysis, both projected paths in* (5) *are non-degenerate. There exist constants $a_p^c, a_p^s > 0$ such that*

$$I_{\mathrm{tr}}\big(E_c^{(m)}; E_s^{(m)}\big) \ge a_p^c \lambda_{p,c}^{(m)2} + a_p^s \lambda_{p,s}^{(m)2}. \tag{14}$$

*The same sensitivity bounds hold along the interpolation between the training and test nuisance distributions.*

Assumption 3 is an explicit structural condition rather than a generic consequence of mutual-information minimization: common–specific independence alone cannot prevent a nuisance factor from entering only one encoder. It therefore restricts attention to the two-channel regime in which the same pseudo-shared factor excites both common-aligned and specific-aligned directions. A linear-Gaussian setting in which (14) holds with explicit constants, together with specializations of the corruption model to the noise and missing regimes, is provided in the supplementary material. Let

$$a_p^{\min} = \min\{a_p^c, a_p^s\}.$$

**Assumption 4** (Shortcut-aligned corruption shift)**.** *Let $Q_p^{\mathrm{tr}}$ and $Q_p^{\mathrm{te}}$ denote the joint distributions of the pseudo-shared factors in the two environments, and define*

$$D_p = W_2\left(Q_p^{\mathrm{tr}}, Q_p^{\mathrm{te}}\right), \tag{15}$$

*where $W_2$ denotes the 2-Wasserstein distance. For the learned corruption-augmented solution, the change in the pseudo-shared factor increases risk by at least*

$$R_{\mathrm{te}}(\widehat{f}_{\mathrm{aug}}) - R_{\mathrm{tr}}(\widehat{f}_{\mathrm{aug}}) \ge b_p D_p - \Omega_\xi, \tag{16}$$

*where $b_p > 0$ measures shortcut use and its alignment with the test shift, and $\Omega_\xi$ bounds the effect of changes in the remaining independent corruption variables.*

Assumption 4 does not state that every unstructured model must use a shortcut. Instead, it characterizes the regime in which the learned baseline has exploited a pseudo-shared training correlation that becomes unreliable at test time. This property can be tested through controlled interventions on $z_p$, as in Section 3.5.

### 3.4 Generalization Analysis

#### 3.4.1 Capacity Control by Structural Constraints

Let $\ell \circ \mathcal{H}$ denote the loss class induced by a reconstruction class $\mathcal{H}$.

**Theorem 1** (Capacity control)**.** *Under Assumption 1,*

$$\widehat{\mathfrak{R}}_n\left(\ell \circ \mathcal{H}_{\mathrm{DAR}}\right) \le \widehat{\mathfrak{R}}_n\left(\ell \circ \mathcal{H}_{\mathrm{aug}}\right). \tag{17}$$

*Furthermore, with probability at least* $1-\delta$*, every* $f \in \mathcal{H}_{\mathrm{DAR}}$ *satisfies*

$$R_{\mathrm{tr}}(f) \le \widehat{R}_n(f) + 2\widehat{\mathfrak{R}}_n\left(\ell \circ \mathcal{H}_{\mathrm{DAR}}\right) + 3M_\ell\sqrt{\frac{\log(2/\delta)}{2n}}. \quad (18)$$

*Proof.* The class inclusion in (11) implies $\ell\circ\mathcal{H}_{\mathrm{DAR}} \subseteq \ell\circ\mathcal{H}_{\mathrm{aug}}$. Rademacher complexity is monotone under set inclusion, which gives (17). Eq. (18) then follows from the standard Rademacher generalization inequality for bounded losses [39]. □

**Interpretation.** Theorem 1 does not claim that the constraints always yield a strictly smaller realized complexity; it establishes the more fundamental fact that the DAR structure cannot enlarge the hypothesis class. DAR thus trades a possible increase in empirical fitting error $\widehat{\Delta}_n$ for a class with no larger estimation complexity.

For subsequent results, define

$$\mathcal{E}_n(\mathcal{H},\delta) = 2\widehat{\mathfrak{R}}_n(\ell\circ\mathcal{H}) + 3M_\ell\sqrt{\frac{\log(4/\delta)}{2n}}. \quad (19)$$

#### 3.4.2 Sensitivity to Pseudo-Shared Shortcuts

The second result characterizes the mechanism by which DAR reduces sensitivity to pseudo-shared factors.

**Theorem 2** (Two-channel shortcut-sensitivity bound)**.** *Under Assumptions 1–3, every* $f \in \mathcal{H}_{\mathrm{DAR}}$ *satisfies*

$$\sum_{m\in\mathcal{M}}\left(\lambda_{p,c}^{(m)2} + \lambda_{p,s}^{(m)2}\right) \le \frac{|\mathcal{M}|\varepsilon_d}{a_p^{\min}}. \quad (20)$$

*If the decoder* $\mathcal{D}$ *is* $L_D$*-Lipschitz, the sensitivity of the reconstructed output to the pseudo-shared factor is bounded by*

$$\mathcal{S}_p(f) \le \underbrace{L_D|\mathcal{M}|\sqrt{\frac{2\varepsilon_d}{a_p^{\min}}}}_{\Lambda_{\mathrm{DAR}}} + \eta_s, \quad (21)$$

*where*

$$\mathcal{S}_p(f) = \left(\mathbb{E}\left\|\partial_{z_p} f(\widetilde{x})\right\|_F^2\right)^{1/2}.$$

*Consequently,*

$$R_{\mathrm{te}}(f) - R_{\mathrm{tr}}(f) \le L_\ell\Lambda_{\mathrm{DAR}}D_p + \Omega_\xi. \quad (22)$$

*Proof.* Combining the DAR constraint $I_{\mathrm{tr}}(E_c^{(m)}; E_s^{(m)}) \le \varepsilon_d$ with Assumption 3, and using $a_p^{\min} \le \min\{a_p^c, a_p^s\}$, gives

$$a_p^{\min}\left(\lambda_{p,c}^{(m)2} + \lambda_{p,s}^{(m)2}\right) \le a_p^c\lambda_{p,c}^{(m)2} + a_p^s\lambda_{p,s}^{(m)2} \le \varepsilon_d$$

for every modality. Summing over $m$ yields (20).

The decoder receives both encoder outputs from all modalities. Applying the chain rule and the Lipschitz bound on $\mathcal{D}$, and adding the residual $\eta_s$ of Assumption 2 for the imperfectly identified stable common subspace, gives

$$\mathcal{S}_p(f) \le L_D\sum_{m\in\mathcal{M}}\left(\lambda_{p,c}^{(m)} + \lambda_{p,s}^{(m)}\right) + \eta_s.$$

The sum contains $2|\mathcal{M}|$ non-negative terms, so Cauchy–Schwarz followed by (20) bounds it as

$$\sum_{m\in\mathcal{M}}\left(\lambda_{p,c}^{(m)} + \lambda_{p,s}^{(m)}\right) \le \sqrt{2|\mathcal{M}|}\left(\sum_{m\in\mathcal{M}}\left(\lambda_{p,c}^{(m)2} + \lambda_{p,s}^{(m)2}\right)\right)^{1/2} \le |\mathcal{M}|\sqrt{\frac{2\varepsilon_d}{a_p^{\min}}},$$

which gives (21). Finally, coupling the training and test pseudo-shared distributions and applying the Lipschitz property of the loss bounds their risk difference by $L_\ell\mathcal{S}_p(f)D_p$. Changes in the remaining corruption variables contribute at most $\Omega_\xi$, yielding (22). □

**Interpretation.** A smaller decoupling budget $\varepsilon_d$ reduces the amount of pseudo-shared information that can jointly occupy the common and specific channels, a larger two-channel margin $a_p^{\min}$ makes the nuisance easier to detect through common–specific dependence, and $\eta_s$ accounts for imperfect identification of the stable common subspace. Importantly, the theorem bounds the sensitivity of the *learned representation* rather than claiming that the larger unstructured class has a worse population optimum: the advantage arises when the empirical unstructured solution relies on a correlation that the test environment changes.

#### 3.4.3 Sufficient Condition for Improved Test Risk

We now combine the empirical fitting gap, estimation complexity, and shortcut-shift terms.

**Theorem 3** (DAR versus corruption-augmented learning)**.** *Let* $\widehat{f}_{\mathrm{DAR}}$ *and* $\widehat{f}_{\mathrm{aug}}$ *be the empirical minimizers defined above. Under Assumptions 1–4, with probability at least* $1-\delta$*,*

$$R_{\mathrm{te}}(\widehat{f}_{\mathrm{DAR}}) - R_{\mathrm{te}}(\widehat{f}_{\mathrm{aug}}) \le \widehat{\Delta}_n + \mathcal{E}_n\left(\mathcal{H}_{\mathrm{DAR}},\delta\right) + \mathcal{E}_n\left(\mathcal{H}_{\mathrm{aug}},\delta\right) - \left(b_p - L_\ell\Lambda_{\mathrm{DAR}}\right)D_p + 2\Omega_\xi. \quad (23)$$

*Therefore, DAR achieves strictly lower test risk whenever*

$$\left(b_p - L_\ell\Lambda_{\mathrm{DAR}}\right)D_p > \widehat{\Delta}_n + \mathcal{E}_n\left(\mathcal{H}_{\mathrm{DAR}},\delta\right) + \mathcal{E}_n\left(\mathcal{H}_{\mathrm{aug}},\delta\right) + 2\Omega_\xi. \quad (24)$$

*Proof.* Add and subtract the two training-environment risks:

$$R_{\mathrm{te}}(\widehat{f}_{\mathrm{DAR}}) - R_{\mathrm{te}}(\widehat{f}_{\mathrm{aug}}) = R_{\mathrm{tr}}(\widehat{f}_{\mathrm{DAR}}) - R_{\mathrm{tr}}(\widehat{f}_{\mathrm{aug}}) + \left[R_{\mathrm{te}}(\widehat{f}_{\mathrm{DAR}}) - R_{\mathrm{tr}}(\widehat{f}_{\mathrm{DAR}})\right] - \left[R_{\mathrm{te}}(\widehat{f}_{\mathrm{aug}}) - R_{\mathrm{tr}}(\widehat{f}_{\mathrm{aug}})\right].$$

Uniform convergence over the two classes and a union bound give

$$R_{\mathrm{tr}}(\widehat{f}_{\mathrm{DAR}}) - R_{\mathrm{tr}}(\widehat{f}_{\mathrm{aug}}) \le \widehat{\Delta}_n + \mathcal{E}_n(\mathcal{H}_{\mathrm{DAR}},\delta) + \mathcal{E}_n(\mathcal{H}_{\mathrm{aug}},\delta).$$

Theorem 2 upper-bounds the shift penalty of DAR by $L_\ell\Lambda_{\mathrm{DAR}}D_p + \Omega_\xi$, whereas Assumption 4 lower-bounds the shift penalty of the learned augmented baseline by $b_pD_p - \Omega_\xi$. Combining the three inequalities gives (23). Condition (24) makes its right-hand side strictly negative. □

**Interpretation.** Theorem 3 exposes the complete trade-off: DAR may sacrifice training fit through $\widehat{\Delta}_n$, but pays no larger statistical complexity and has lower pseudo-shared sensitivity when $\Lambda_{\mathrm{DAR}}$ is small, so its advantage becomes visible once the shortcut penalty avoided at test time exceeds the fitting, estimation, and residual-corruption terms.

**Corollary 1** (Increasing corruption shift)**.** *Fix the learned models, sample size, and decoupling parameters. If $b_p > L_\ell \Lambda_{\mathrm{DAR}}$, then the right-hand side of* (23) *decreases as $D_p$ increases. Consequently, once the corruption shift exceeds*

$$D_p > \frac{\widehat{\Delta}_n + \mathcal{E}_n(\mathcal{H}_{\mathrm{DAR}}, \delta) + \mathcal{E}_n(\mathcal{H}_{\mathrm{aug}}, \delta) + 2\Omega_\xi}{b_p - L_\ell \Lambda_{\mathrm{DAR}}}, \tag{25}$$

*DAR has lower test risk with probability at least $1 - \delta$.*

This result provides a theoretical interpretation of the empirical trend that DAR's advantage is often more pronounced under severe corruption: stronger corruption can increase the instability of the shortcut used by an unstructured model, while DAR's leakage budget remains controlled by $\varepsilon_d$ and $a_p^{\min}$.

*Remark* 1 (Unified low-quality observations). The analysis applies to both missingness and additive noise through the operator in (3). For missing observations, $D_p$ can represent a change in content-dependent mask correlations, while $\Omega_\xi$ captures independent changes in the missing rate. For noisy observations, $D_p$ describes changes in cross-modal environmental or sensor artifacts, and $\Omega_\xi$ captures changes in independent noise magnitude. The benefit of DAR is therefore associated with the same structural principle in both cases: stable common information is recovered across modalities, whereas modality-specific information is recovered within its own modality.

### 3.5 Synthetic Validation of Observable Predictions

The quantities appearing above — semantic factors, pseudo-shared factors, and corruption shift — cannot be observed directly in real datasets. We therefore construct a controlled synthetic benchmark in which they are known by design, with the aim of testing the *observable* predictions of the analysis rather than numerically estimating the Rademacher-complexity terms or theorem constants.

**Synthetic generative process.** We instantiate (1) with three modalities $\{x_l, x_v, x_a\}$ and omit the temporal dimension. The common factor $z_c$ is sampled from a Gaussian mixture, and each modality-specific factor $z_s^{(m)}$ is sampled independently from a modality-dependent Gaussian mixture and passed through $\tanh$. During training the pseudo-shared factors are strongly correlated across modalities, $z_p^{(m,\mathrm{tr})} = z_p^{\mathrm{base}} + \rho\nu_m$; at test time this correlation is deliberately altered by sign-flipping the nuisance in one modality and cyclically permuting it in another, producing a controlled corruption shift with $D_p > 0$ while leaving the semantic factors unchanged. Corruption is applied as $\widetilde{x}_m = x_m \odot \mathrm{mask}_m$, where the probability of masking a feature depends on $\|z_s^{(m)}\|$ and contiguous dimensions are masked together, so that the observations combine missing-not-at-random and block missingness with a pseudo-shared shortcut whose cross-modal behavior changes at test time.


Fig. 2. Test reconstruction error on missing entries. Decouple achieves lower and more stable $\mathrm{MSE}_{\mathrm{miss}}$ than Direct under the controlled pseudo-shared corruption shift.

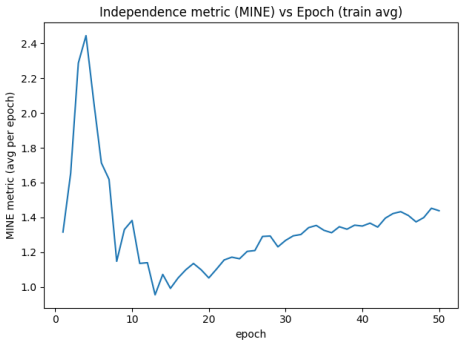



Fig. 3. MINE-based estimate of common–specific dependence during training. The estimate is used as a diagnostic proxy rather than a direct numerical estimate of the theoretical leakage bound.

**Compared methods and evaluation. Direct** concatenates the corrupted modalities and their masks and uses an MLP to reconstruct the clean semantic features. **Decouple** implements the DAR principle by encoding every modality into common and specific channels, aligning common representations across modalities, reducing common–specific dependence, and restoring the two components through structurally distinct heads; the complete architecture is introduced in Section 4. The primary metric is the reconstruction error computed only on missing entries,

$$\mathrm{MSE}_{\mathrm{miss}} = \frac{\sum_m \sum_i (1 - \mathrm{mask}_{m,i}) \|\widehat{x}_{m,i} - x^\star_{m,i}\|_2^2}{\sum_m \sum_i (1 - \mathrm{mask}_{m,i})}, \tag{26}$$

and we additionally track a MINE-based estimate of $I(\widetilde{\mathbf{H}}^{\mathrm{com}}; \widetilde{\mathbf{H}}^{\mathrm{spec}})$ during training as a diagnostic proxy for the decoupling budget $\varepsilon_d$.

**Results.** Fig. 2 shows that Decouple attains consistently lower missing-entry reconstruction error than Direct, with a more stable late-stage trajectory. Because the semantic factors are held fixed while only the pseudo-shared correlation is changed, this is precisely the trade-off described by Theorem 3: the structured model pays a modest fitting cost but is markedly less sensitive to the unstable cross-modal shortcut. The MINE estimate in Fig. 3 first increases, as the variational estimator begins to identify statistical dependence, and then decreases as the decoupling objective takes effect. Since

MINE is itself a learned estimator, its absolute value should not be interpreted as the quantity in (14); the declining late-stage trend nevertheless provides diagnostic evidence that common–specific dependence is being suppressed during optimization, which is the mechanism formalized by Theorem 2. These results should be read as controlled empirical evidence for the proposed mechanism rather than as direct numerical verification of the theoretical complexity bounds.

## 4 THE DAR FRAMEWORK

Guided by the analysis of Section 3, we now instantiate the decouple–restore–reconstruct principle as a concrete learning framework, **Decoupled-Adaptive Reconstruction (DAR)**.

**Relation to the preliminary version.** A first version of DAR was introduced in [40], developed for a single setting: sequential multimodal sentiment analysis under random modality missingness, realized by a fixed attention-based network. That pipeline is retained here as one special case, and the components inherited from it — the two-branch encoder, the InfoNCE/CLUB decoupling objective, the two restoration heads, and the gated output module — are described only to the level of detail needed for self-containment, with derivations and implementation details left to [40]. New in this section are (i) a formulation over the unified corruption operator (3), so that missingness and noise are processed by one pipeline instead of two; (ii) a specification of every module by its *information-flow role* — which representations it consumes and produces, and which constraint it must satisfy — rather than by a network type, which is what allows the DAR hypothesis class of Definition 2 to be realized by different architectures; and (iii) two reference instantiations of these roles (Section 4.5), a *sequential* one for the sequence benchmarks and a *static feed-forward* one for the static-feature benchmarks. Self-attention is therefore one possible realization of an aggregation role, not a defining ingredient of DAR.

Fig. 4 summarizes the resulting pipeline: a unified low-quality input (Section 4.1) is decoupled following the structural prior of Theorem 2 (Section 4.2), restored by two heads that exploit different information sources (Section 4.3), and reconstructed by gated fusion (Section 4.4).

### 4.1 Problem Setup and Unified Corruption

Let $\mathcal{M}$ denote the set of modalities and $y$ the target, either categorical ($y \in \mathbb{R}^C$) or continuous ($y \in \mathbb{R}$). Two input regimes share the same pipeline. For *sequential inputs*, each modality is a temporal sequence $\mathbf{U}_m \in \mathbb{R}^{T_m \times d_m}$, as in the language, visual, and acoustic streams of multimodal sentiment analysis ($\mathcal{M} = \{l, v, a\}$); for *static inputs*, each modality is a single feature vector $\mathbf{U}_m \in \mathbb{R}^{d_m}$ from a pre-trained unimodal backbone, as in image–text classification or RGB-D scene recognition ($\mathcal{M} = \{\mathrm{rgb}, \mathrm{depth}\}$). The static regime is the $T_m = 1$ special case of the sequential one, so we keep the sequence notation throughout and read every time-dimension operation as the identity when $T_m = 1$.

Following the unified corruption operator of (3), the low-quality input $\widetilde{\mathbf{U}}_m$ is produced by one of two training-time scenarios that share the same downstream pipeline. Under *missing-corruption*, a proportion $k \in [0, 0.9]$ of the entries of each modality is randomly masked; numerical features are zero-filled and language masks are filled with $[\mathrm{UNK}]$ tokens. Under *noise-corruption*, zero-mean Gaussian noise of standard deviation $\sigma$ is added to numerical features, while text undergoes blank-token replacement. Both scenarios preserve the shape of $\mathbf{U}_m$; the clean $\mathbf{U}_m$ is used *only* for supervision during training, and at inference only $\widetilde{\mathbf{U}}_m$ is available. A modality alignment layer — a 1D-convolution followed by a non-linear projection for sequential inputs, a non-linear projection for static inputs — then maps all modalities to a common shape, yielding $\mathbf{U}_m^1, \widetilde{\mathbf{U}}_m^1 \in \mathbb{R}^{t \times d}$.

### 4.2 Stage 1: Decouple Module

Two parallel encoders factorize each modality into a modal-common component and a modal-specific component,

$$\widetilde{\mathbf{H}}_m^{\mathrm{com}} = E_c\big(\widetilde{\mathbf{U}}_m^1; \theta_m^c\big), \quad \widetilde{\mathbf{H}}_m^{\mathrm{spec}} = E_s\big(\widetilde{\mathbf{U}}_m^1; \theta_m^s\big), \tag{27}$$

and the same encoders applied to $\mathbf{U}_m^1$ produce the supervisory targets $\mathbf{H}_m^{\mathrm{com}}, \mathbf{H}_m^{\mathrm{spec}}$. Following Theorem 2, the factorization should drive $I(\widetilde{\mathbf{H}}_m^{\mathrm{com}}; \widetilde{\mathbf{H}}_m^{\mathrm{spec}}) \to 0$ while maximizing cross-modal common similarity and within-modal specific similarity, which we implement through the composite objective

$$\mathcal{L}_{\mathrm{decouple}} = \lambda(\mathcal{L}_{\mathrm{sim}} + \mathcal{L}_{\mathrm{diff}}) + \mathcal{L}_{\mathrm{re}}. \tag{28}$$

**Mutual-information bounds.** As in [40], the two mutual-information terms are estimated by a variational lower and upper bound respectively. Writing $s(\mathbf{z}, \mathbf{z}') = \exp(\mathrm{sim}(\mathbf{z}, \mathbf{z}')/\tau)$ for the cosine similarity with temperature $\tau$, and taking the positive set $\mathcal{P}$ and negative set $\mathcal{N}$ of Fig. 5, the similarity term is the InfoNCE [38] lower bound

$$\mathcal{L}_{\mathrm{sim}} = -\frac{1}{|\mathcal{P}|} \sum_{(\mathbf{z}_1, \mathbf{z}_2) \in \mathcal{P}} \log \frac{s(\mathbf{z}_1, \mathbf{z}_2)}{\sum_{(\mathbf{z}_1, \mathbf{z}_i) \in \mathcal{N}} s(\mathbf{z}_1, \mathbf{z}_i)}, \tag{29}$$

computed separately for the two factorizations of Fig. 5 and summed, $\mathcal{L}_{\mathrm{sim}} = \mathcal{L}_{\mathrm{sim}}^{\mathrm{com}} + \mathcal{L}_{\mathrm{sim}}^{\mathrm{spec}}$. Sequence inputs are pooled along the time axis into a single vector $\mathbf{z} \in \mathbb{R}^d$ before the similarity is evaluated; for static inputs ($t = 1$) this pooling is the identity. The dissimilarity term minimizes $I(\widetilde{\mathbf{H}}_m^{\mathrm{com}}; \widetilde{\mathbf{H}}_m^{\mathrm{spec}})$ through the CLUB upper bound [15],

$$\mathcal{L}_{\mathrm{diff}} = \frac{1}{N} \sum_{i=1}^{N} \log q_\theta\big(\widetilde{\mathbf{H}}_{m,i}^{\mathrm{com}} | \widetilde{\mathbf{H}}_{m,i}^{\mathrm{spec}}\big) - \frac{1}{N^2} \sum_{i,j=1}^{N} \log q_\theta\big(\widetilde{\mathbf{H}}_{m,i}^{\mathrm{com}} | \widetilde{\mathbf{H}}_{m,j}^{\mathrm{spec}}\big), \tag{30}$$

where the variational network $q_\theta$ approximates $p(\widetilde{\mathbf{H}}_m^{\mathrm{com}} \mid \widetilde{\mathbf{H}}_m^{\mathrm{spec}})$ and is trained alternately with the main network. Its parameterization is a role, not a fixed choice, and follows Table 1.

**Self-regression anchor.** The MI constraints alone admit degenerate solutions that discard information about $\mathbf{U}_m^1$. We therefore anchor the decomposition by regressing each modality from the *other* modalities' common features together with its own specific feature,

$$\mathcal{L}_{\mathrm{re}}^m = \Big\| \widetilde{\mathbf{U}}_m^1 - \mathcal{D}_m\Big(\mathrm{Concat}\big(\{\widetilde{\mathbf{H}}_{m'}^{\mathrm{com}}\}_{m' \neq m}, \widetilde{\mathbf{H}}_m^{\mathrm{spec}}\big)\Big) \Big\|_F^2, \tag{31}$$

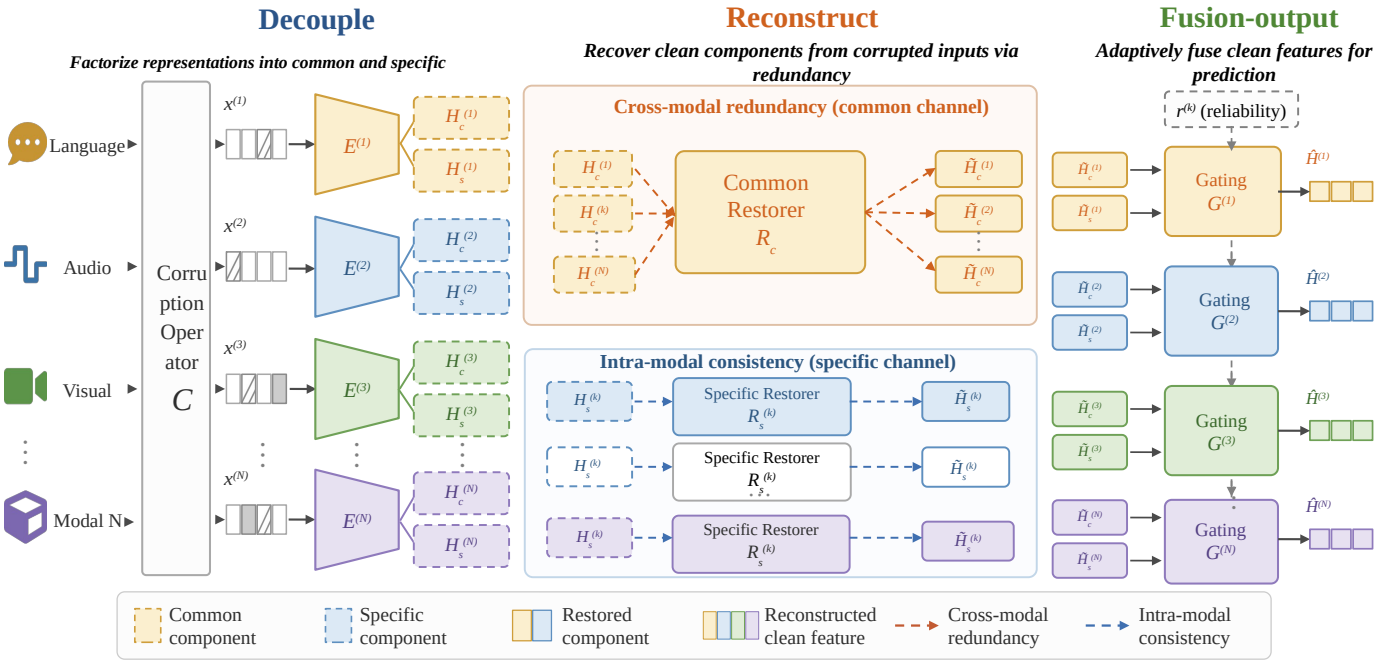



Fig. 4. Overall architecture of DAR on $N$ modalities. The corruption operator $C$ of (3) maps every modality to a low-quality input $x^{(k)}$: blank cells are intact entries, hatched cells are erased entries, and solid gray cells are noise-perturbed entries. Stage 1 splits $x^{(k)}$ into a modal-common feature $H_c^{(k)}$ (dashed orange) and a modal-specific feature $H_s^{(k)}$ (dashed blue); stage 2 repairs the two channels separately, the common restorer $R_c$ drawing on *all* modalities and each specific restorer $R_s^{(k)}$ on its own modality alone; stage 3 applies the reliability-driven gate $G^{(k)}$ and feeds the fused feature $\hat{H}^{(k)}$ to the prediction head. Table 1 lists the two realizations of each role-defined module. The figure writes $H_c^{(k)}, H_s^{(k)}, \tilde{H}_c^{(k)}, \tilde{H}_s^{(k)}$, and $\hat{H}^{(k)}$ for $\widetilde{\mathbf{H}}_m^{\mathrm{com}}, \widetilde{\mathbf{H}}_m^{\mathrm{spec}}, \widehat{\mathbf{H}}_m^{\mathrm{com}}, \widehat{\mathbf{H}}_m^{\mathrm{spec}}$, and $\mathbf{H}_{\mathrm{fused},m}$ of the text.

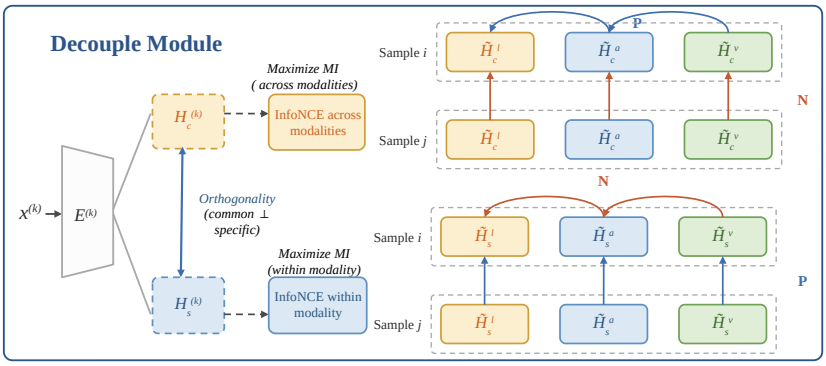



Fig. 5. The decouple module and its positive (P) / negative (N) pairs, shown for $\mathcal{M} = \{l, a, v\}$ and two instances $i, j$ of a batch. For the common channel, positives are common features of different modalities from the same instance and negatives come from different instances; for the specific channel, positives are specific features of the same modality across instances and negatives come from different modalities. Both panels operate on the corrupted branch.

with $\mathcal{L}_{\mathrm{re}} = \sum_{m\in\mathcal{M}} \mathcal{L}_{\mathrm{re}}^m$. This directly encodes the redundancy property of the generative model of Section 3.1: what is shared must be recoverable from the other modalities, and what is not must be carried by the specific channel.

### 4.3 Stage 2: Restore Module

The restore module predicts what the decoupled representations of a corrupted input *would have been* on the clean input. The two channels have different information sources — common features are shared and therefore remain partially observable in other modalities, whereas specific features are private and can only be recovered from intra-modal structure — so DAR uses two structurally distinct heads and a private decoder $\mathcal{D}_m$ back to the signal space:

$$\widehat{\mathbf{H}}_m^{\mathrm{com}} = E_{\mathrm{com}}^m\Big(\mathrm{Concat}\big(\{\widetilde{\mathbf{H}}_{m'}^{\mathrm{com}}\}_{m'\in\mathcal{M}}\big); \theta_{\mathrm{com}}^m\Big), \tag{32}$$

$$\widehat{\mathbf{H}}_m^{\mathrm{spec}} = E_{\mathrm{spec}}^m\big(\widetilde{\mathbf{H}}_m^{\mathrm{spec}}; \theta_{\mathrm{spec}}^m\big), \tag{33}$$

$$\widehat{\mathbf{U}}_m^1 = \mathcal{D}_m\big(\mathrm{Concat}(\widehat{\mathbf{H}}_m^{\mathrm{com}}, \widehat{\mathbf{H}}_m^{\mathrm{spec}})\big). \tag{34}$$

Because (32) aggregates every modality, the common channel stays well-posed even when each individual $\widetilde{\mathbf{H}}_m^{\mathrm{com}}$ is corrupted. Restoration is supervised by the clean-feature targets,

$$\mathcal{L}_{\mathrm{recon}} = \big\|\widehat{\mathbf{H}} - \mathbf{H}\big\|_F^2 + \big\|\widehat{\mathbf{U}}^1 - \mathbf{U}^1\big\|_F^2, \tag{35}$$

where $\hat{\mathbf{H}} = (\hat{\mathbf{H}}^{\mathrm{com}}, \hat{\mathbf{H}}^{\mathrm{spec}})$ and $\mathbf{H} = (\mathbf{H}^{\mathrm{com}}, \mathbf{H}^{\mathrm{spec}})$ are stacked across all modalities.

### 4.4 Stage 3: Fusion-Output Module

**Adaptive gated reconstruction.** Since the restored representation may itself be imperfect, it is combined with the directly encoded features through a learned gate,

$$\mathbf{g} = \sigma\big(\mathbf{W}_g[\hat{\mathbf{H}}, \widetilde{\mathbf{H}}] + \mathbf{b}_g\big), \tag{36}$$

$$\mathbf{H}_{\mathrm{fused}} = \mathbf{g} \odot \hat{\mathbf{H}} + (1 - \mathbf{g}) \odot \widetilde{\mathbf{H}}, \tag{37}$$

so that heavily corrupted positions rely on $\hat{\mathbf{H}}$ and lightly corrupted ones retain $\widetilde{\mathbf{H}}$ — the source of the *adaptive* in the framework's name.

**Role-defined aggregation.** The fused features are then summarized by two role-defined operators whose realization depends on the input regime and is deferred to Section 4.5. The *common aggregator* $\mathcal{A}_{\mathrm{com}}$ exploits the fact that fused common features share similar distributions across modalities and summarizes them jointly, whereas the *specific interaction operator* $\mathcal{A}^m_{\mathrm{spec}}$ lets each modality's specific feature, whose distribution differs markedly across modalities, be enhanced by those of the remaining modalities:

$$\mathbf{h}^{\mathrm{com}} = \mathcal{A}_{\mathrm{com}}\big(\{\mathbf{H}^{\mathrm{com}}_{\mathrm{fused},m}\}_{m\in\mathcal{M}}\big), \tag{38}$$

$$\mathbf{h}^{\mathrm{spec}}_m = \mathcal{A}^m_{\mathrm{spec}}\big(\mathbf{H}^{\mathrm{spec}}_{\mathrm{fused},m}; \{\mathbf{H}^{\mathrm{spec}}_{\mathrm{fused},m'}\}_{m'\neq m}\big), \tag{39}$$

with $\mathbf{h}^{\mathrm{spec}} = \mathrm{Concat}(\{\mathbf{h}^{\mathrm{spec}}_m\}_{m\in\mathcal{M}})$.

**Prediction and bootstrap supervision.** The prediction is $\hat{y} = \mathrm{MLP}(\mathrm{Concat}(\mathbf{h}^{\mathrm{com}}, \mathbf{h}^{\mathrm{spec}}))$. To counter the objective interference discussed in Section 1, the same head is reused on the clean-input features, $\hat{y}_{\mathrm{boot}} = \mathrm{MLP}(\mathbf{H})$, and both branches are supervised, $\mathcal{L}_{\mathrm{task}} = \mathrm{Loss}(y, \hat{y}) + \mathrm{Loss}(y, \hat{y}_{\mathrm{boot}})$. The bootstrap branch forces the encoder to keep features that are useful for the task even when no restoration is needed, blocking the trivial solution of equating corrupted and clean features at the expense of expressivity. The overall objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{task}} + \alpha\,\mathcal{L}_{\mathrm{decouple}} + \beta\,\mathcal{L}_{\mathrm{recon}}, \tag{40}$$

with $\alpha, \beta > 0$, and is identical for every instantiation of the framework.

### 4.5 Instantiating the Role-Defined Modules

The stages above constrain *what* each module consumes, produces, and optimizes, but not *how* it is parameterized. We use the two reference instantiations of Table 1. Both optimize the identical objective (40); they differ only in the network family realizing the alignment layer, the variational network $q_\theta$, and the aggregation operators $\mathcal{A}_{\mathrm{com}}, \mathcal{A}_{\mathrm{spec}}$.

**Sequential instantiation.** For temporal inputs (MOSI, MOSEI, SIMS) we adopt the sequential realization of [40]: $\mathcal{A}_{\mathrm{com}}$ is a multi-layer self-attention block [41] applied along the temporal dimension to the fused common features of all modalities, with the last frame of the final layer taken as the holistic vector $\mathbf{h}^{\mathrm{com}}$; $\mathcal{A}^m_{\mathrm{spec}}$ is a set of pairwise crossmodal attention units [17] in which the specific feature of modality $m$ acts as the query and that of each remaining modality as key and value, the resulting enhanced features being concatenated; and $q_\theta$ in (30) is a bidirectional LSTM [42] followed by a non-linear projection, matching the temporal structure of the hidden representations.

TABLE 1
Two reference instantiations of the role-defined modules. All losses, the gated reconstruction, and the restoration heads are shared; only the modules below differ.

| Module (role) | Sequential | Static (feedforward) |
|---|---|---|
| Alignment layer | 1D-Conv + projection | non-linear projection |
| MI variational net $q_\theta$ | BiLSTM + projection | feedforward network |
| Common aggregator $\mathcal{A}_{\mathrm{com}}$ | temporal self-attention | feedforward network |
| Specific interaction $\mathcal{A}_{\mathrm{spec}}$ | pairwise cross-attention | feedforward network |
| Sequence pooling (InfoNCE) | time averaging | identity ($t = 1$) |

**Algorithm 1** Training of DAR

**Require:** Training data $\{(\mathbf{U}^{(i)}_m, y^{(i)})\}_{i=1}^N$; corruption level ($k$ or $\sigma$); hyperparameters $\lambda, \alpha, \beta, \tau$; instantiation (Table 1).

1: Generate corrupted views $\widetilde{\mathbf{U}}^{(i)}_m$ via (3); initialize all modules according to the chosen instantiation.
2: **for** each minibatch **do**
3: Encode $\widetilde{\mathbf{H}}^{\mathrm{com}}_m, \widetilde{\mathbf{H}}^{\mathrm{spec}}_m$ and the clean targets $\mathbf{H}^{\mathrm{com}}_m, \mathbf{H}^{\mathrm{spec}}_m$ via $E_c, E_s$.
4: Update $q_\theta$ by maximizing $\log q_\theta(\widetilde{\mathbf{H}}^{\mathrm{com}}_m \mid \widetilde{\mathbf{H}}^{\mathrm{spec}}_m)$.
5: Compute $\mathcal{L}_{\mathrm{decouple}}$ from (29)–(31).
6: Restore $\hat{\mathbf{H}}^{\mathrm{com}}_m, \hat{\mathbf{H}}^{\mathrm{spec}}_m, \hat{\mathbf{U}}^1_m$ via (32)–(34) and compute $\mathcal{L}_{\mathrm{recon}}$ via (35).
7: Gate, aggregate via (38)–(39), and predict $\hat{y}, \hat{y}_{\mathrm{boot}}$.
8: Update all parameters except $q_\theta$ with $\mathcal{L}_{\mathrm{total}}$ of (40).
9: **end for**
10: **Output:** trained DAR model.

**Static feedforward instantiation.** When each modality is a single feature vector ($t = 1$; MVSA, NYU Depth V2, SUN RGB-D), temporal machinery is vacuous: self-attention over a length-one sequence degenerates to a per-token linear map, and recurrence has no time axis to traverse. We therefore realize *every* sequence-specific module as a plain multi-layer perceptron. $\mathcal{A}_{\mathrm{com}}$ becomes an MLP on the concatenation of the fused common features of all modalities; $\mathcal{A}^m_{\mathrm{spec}}$ becomes a per-modality MLP on the concatenation of $\mathbf{H}^{\mathrm{spec}}_{\mathrm{fused},m}$ with the specific features of the other modalities; $q_\theta$ becomes a feedforward conditional density network; and the encoders, restorers, and decoders operate directly on the aligned feature vectors. All losses in (40) are computed exactly as in the sequential case.

**Design rationale.** Beyond matching the inductive bias to the data, the two instantiations substantiate the theory: Definition 2 constrains only the existence of modality-wise common/specific encoders and their mutual information and never references attention, so any Lipschitz realization inherits Theorems 1–3. Consistent gains from both (Section 5) therefore indicate that robustness follows the decouple–restore–reconstruct principle rather than a particular fusion architecture. The feedforward variant is also light enough for DAR to act as a plug-in front-end for standard backbones such as Late Fusion and Concat (Section 5.3).

### 4.6 Algorithmic Summary

TABLE 2
Overview of the six evaluation benchmarks. L/V/A denote language, visual, and acoustic streams; I/T denote image and text. The two corruption regimes are evaluated with the two instantiations of Table 1.

| Benchmark | Modalities | Task | Corruption | Inst. |
|---|---|---|---|---|
| MOSI [43] | L, V, A | Sentiment | Erasure, $r \in [0, 0.9]$ | Seq. |
| MOSEI [44] | L, V, A | Sentiment | Erasure, $r \in [0, 0.9]$ | Seq. |
| SIMS [45] | L, V, A | Sentiment (Chinese) | Erasure, $r \in [0, 0.9]$ | Seq. |
| MVSA-Single [46] | I, T | Sentiment | Gaussian, $\sigma \in \{0, 0.5, 1.0\}$ | FF |
| NYU Depth V2 [47] | RGB, D | Scene rec. (10) | Gaussian, $\sigma \in \{0, 0.5, 1.0\}$ | FF |
| SUN RGB-D [48] | RGB, D | Scene rec. (19) | Gaussian, $\sigma \in \{0, 0.5, 1.0\}$ | FF |

Algorithm 1 summarizes the training procedure, which is shared by both instantiations. At inference, only the corrupted input $\widetilde{\mathbf{U}}_m$ is required: restoration and gated fusion produce $\hat{y}$ end-to-end.

## 5 Experiments

Section 3.5 has already validated the observable predictions of the analysis on a controlled synthetic benchmark, where the semantic factors, the pseudo-shared factors, and the corruption shift are known by construction. We now turn to real data. The experiments are organized around the central claim of this work — that a single decouple–restore–reconstruct pipeline handles missingness and noise alike — and are therefore reported regime by regime rather than dataset by dataset. Section 5.1 fixes the protocol shared by both regimes. Sections 5.2 and 5.3 then report the two main comparisons: erasure corruption on three sequential multimodal sentiment benchmarks, and additive Gaussian corruption on one image–text and two RGB-D benchmarks. Because the first uses the attention–recurrent instantiation and the second the static feedforward instantiation of Section 4.5, reading them back to back also tests whether robustness survives a change of architecture. Section 5.4 then attributes the gains to individual loss terms in *both* regimes and asks what the two attributions have in common, and Section 5.5 examines the learned representations directly. Together these results provide evidence consistent with the unified treatment of low-quality data predicted by Theorem 3 and Remark 1.

### 5.1 Experimental Setup

Table 2 summarizes the six benchmarks, the corruption applied to each, and the DAR instantiation used. Protocol details common to both regimes are given below; regime-specific details follow.

#### 5.1.1 Sequential Benchmarks and Erasure Corruption

**Datasets. MOSI** [43] contains 2,199 multimodal samples (1,284 / 229 / 686 train/val/test) labeled with continuous sentiment scores in $[-3, 3]$. **MOSEI** [44] contains 22,856 video clips (16,326 / 1,871 / 4,659) with the same sentiment annotation. **SIMS** [45] additionally provides Chinese-language sentiment annotations and is used for cross-language validation.

**Feature extraction and erasure.** Following [49], language is encoded with BERT [50], audio with Librosa [51], and visual with OpenFace [52]. Visual and acoustic erasures are zero-filled; language erasures use $[\mathrm{UNK}]$.

**Evaluation protocol.** Following [6], we report the average over 3 random seeds and 10 test-time missing rates from $0.0$ to $0.9$ at $0.1$ intervals. Classification metrics include 7-class, 5-class, and 2-class accuracy (with negative/non-negative and negative/positive variants for Acc-2). Regression metrics are mean absolute error (MAE) and Pearson correlation (Corr). We select the model with the smallest overall loss as optimal, unlike the per-metric best-model selection used in some baselines.

**Baselines.** We compare against MISA [33], Self-MM [21], MMIM [53], CENET [54], TETFN [55], TFR-Net [5], ALMT [49], and the most recent LNLN [6]. Baseline results for MISA, Self-MM, MMIM, CENET, TETFN, and ALMT are reproduced via the unified MMSA [56] framework with default hyperparameters; LNLN is reproduced from open-source code; TFR-Net results are taken from the LNLN paper as they used per-metric best models, providing a fair upper bound.

#### 5.1.2 Static Benchmarks and Gaussian Corruption

**Datasets. NYU Depth V2** [47] and **SUN RGB-D** [48] are RGB-D scene-recognition benchmarks; following [14], we use 10 categories for NYU and 19 for SUN. **MVSA-Single** [46] is an image–text sentiment dataset. In all of these benchmarks each modality is provided as a single static feature vector, matching the static-input regime of Section 4.1.

**Noise injection protocol.** For each test sample, $50\%$ of the modalities are corrupted: image modalities receive zero-mean Gaussian noise of standard deviation $\sigma \in \{0, 0.5, 1.0\}$; text modalities undergo blank-token replacement. This is the same protocol used by QMF [8] and PDF [9], so that the numbers in Tables 5 and 6 are directly comparable to the published ones. The same three noise levels are used throughout the paper, including the representation-level study of Section 5.5; baselines marked "with noise(0.5)" are additionally trained with $\sigma = 0.5$ noise augmentation.

**Baselines and integration.** We compare against Late Fusion, Concat, MMTM [57], TMC [14], QMF [8], DynMM [30], and PDF [9]. The modular design of DAR allows it to serve as a denoising front-end for any fusion strategy. We evaluate two integration variants: **DAR + Late Fusion**, in which the restored unimodal predictions are averaged at the decision level, and **DAR + Concat**, in which the restored unimodal representations are concatenated and mapped to the target via a fully connected layer. Because these benchmarks provide static feature vectors rather than sequences, both variants use the purely feedforward instantiation of Section 4.5: every attention and recurrent component of the sequential realization is replaced by a plain feedforward network, while the loss in Eq. (40) is left unchanged.

#### 5.1.3 Implementation Details

For the sequential instantiation we use the loss weights $\lambda = 0.7$, $\alpha = 0.1$, $\beta = 0.1$ and a training-time missing rate $k = 0.3$ on MOSI and $k = 0.4$ on MOSEI, both selected on validation data (Section 5.4.4). For RGB-D scene recognition we use ResNet-18 [58] pretrained on ImageNet [59]; for image–text we use ResNet-152 for images and BERT [50] for text. All models are trained with Adam at learning rate

TABLE 3
Robustness comparison on MOSI and MOSEI under varying missing rates. Best results are in **bold**, second best are underlined. Acc-2 / F1 are reported as *negative-vs.-non-negative* / *negative-vs.-positive*.

| Method | MOSI | | | | | | MOSEI | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Acc-7 | Acc-5 | Acc-2 | F1 | MAE↓ | Corr | Acc-7 | Acc-5 | Acc-2 | F1 | MAE↓ | Corr |
| MISA | 28.90 | 31.67 | 69.15 / 70.74 | 68.50 / 70.23 | 1.092 | 0.508 | 38.92 | 39.28 | 76.21 / 72.12 | 70.76 / 65.50 | 0.800 | 0.490 |
| Self-MM | 30.78 | 34.03 | 68.75 / 70.89 | 65.47 / 67.90 | 1.070 | 0.518 | 46.40 | 46.78 | 71.18 / 72.75 | 70.45 / 70.99 | 0.695 | 0.498 |
| MMIM | 31.51 | 34.92 | 69.22 / 71.08 | 67.34 / 69.42 | 1.077 | 0.511 | 44.04 | 44.42 | 75.99 / 71.47 | 70.63 / 64.97 | 0.739 | 0.459 |
| CENET | 29.78 | 33.23 | 66.41 / 69.47 | 62.65 / 65.38 | 1.088 | 0.496 | **47.18** | 47.93 | 75.96 / 74.10 | 73.28 / 70.51 | 0.685 | 0.525 |
| TETFN | 29.89 | 33.20 | 68.66 / 70.89 | 65.11 / 67.64 | 1.087 | 0.512 | 46.31 | 47.03 | 71.63 / 71.84 | 68.91 / 68.14 | 0.714 | 0.508 |
| TFR-Net | 29.54 | 34.67 | 68.15 / 66.35 | 61.73 / 60.06 | 1.200 | 0.459 | 46.83 | 34.67 | 73.62 / 77.23 | 68.80 / 71.99 | 0.697 | 0.489 |
| ALMT | 30.35 | 32.92 | 68.27 / 70.55 | 64.47 / 67.07 | 1.083 | 0.506 | 42.01 | 42.58 | 76.75 / 72.96 | 72.00 / 67.16 | 0.754 | 0.511 |
| LNLN | 32.80 | 36.12 | 71.11 / 72.22 | 71.33 / 72.34 | **1.066** | 0.505 | 45.42 | 46.17 | 75.27 / 76.98 | 74.97 / 77.39 | 0.692 | 0.530 |
| **DAR (Ours)** | **34.47** | **38.65** | **71.60 /73.18** | **71.51 /73.15** | 1.069 | **0.520** | 47.01 | **48.02** | **77.48 /78.14** | **77.44 / 77.51** | **0.665** | **0.583** |

TABLE 4
Performance comparison on the SIMS dataset. Acc-$k$ denotes $k$-class accuracy. Best results in **bold**, second best underlined.

| Method | Acc-5 | Acc-3 | Acc-2 | F1 | MAE↓ | Corr |
|---|---|---|---|---|---|---|
| MISA | 23.22 | 45.89 | 56.41 | 54.95 | 0.583 | 0.307 |
| Self-MM | 32.69 | **56.94** | **73.28** | 69.91 | 0.514 | 0.365 |
| MMIM | 29.67 | 52.98 | 69.19 | 69.14 | 0.520 | 0.373 |
| CENET | 20.42 | 49.63 | 69.10 | 59.13 | 0.595 | 0.044 |
| TETFN | 32.28 | 54.29 | 72.12 | 68.88 | 0.507 | 0.381 |
| ALMT | 30.33 | 55.12 | 71.99 | 68.90 | 0.504 | 0.392 |
| LNLN | 32.43 | 55.80 | 69.45 | 69.83 | 0.514 | 0.398 |
| DAR (Ours) | **34.56** | 56.78 | 70.94 | **70.35** | **0.501** | **0.406** |

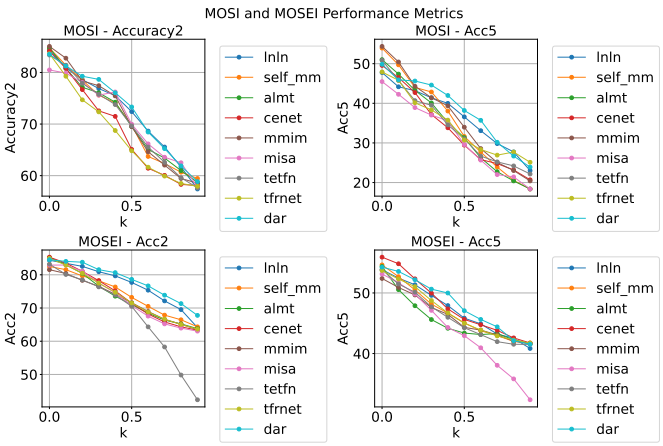



Fig. 6. Acc-2 and Acc-5 of DAR vs. baselines as a function of the test-time missing rate.

$10^{-4}$, and $\lambda, \alpha, \beta$ are kept at the values above so that no per-regime tuning is involved. We report mean and worst-case accuracy over 10 runs (NYU, SUN) and 5 runs (MVSA).

### 5.2 Robustness under Missing Modalities

**Results.** Table 3 reports robustness on MOSI and MOSEI. DAR achieves state-of-the-art on most metrics across both datasets. On MOSEI it attains the best Acc-2 under both binarization schemes, exceeding the strongest baseline on each (77.48 vs. 76.75 for ALMT and 78.14 vs. 77.23 for TFR-Net), and lowers MAE from 0.685 to 0.665; measured

TABLE 5
Robustness on **MVSA-Single** under Gaussian noise ($\sigma$) on 50% of the modalities. Baseline results from [8], [9]. Best results are in **bold**, second best are underlined.

| Method | $\sigma = 0$ | | $\sigma = 0.5$ | | $\sigma = 1.0$ | |
|---|---|---|---|---|---|---|
| | Avg | Worst | Avg | Worst | Avg | Worst |
| Late Fusion | 76.88 | 74.76 | 63.46 | 58.57 | 55.16 | 47.78 |
| Late Fusion (noise=0.5) | 78.04 | 77.26 | 69.76 | 65.31 | 53.63 | 48.74 |
| Concat | 65.59 | 64.74 | 50.70 | 44.70 | 46.12 | 41.81 |
| Concat (noise=0.5) | 64.59 | 63.01 | 55.21 | 47.59 | 46.65 | 40.46 |
| TMC | 74.88 | 71.10 | 66.72 | 60.12 | 60.36 | 53.37 |
| QMF | 78.07 | 76.30 | 73.85 | 71.10 | 61.28 | 57.61 |
| DynMM | 79.07 | 78.23 | 67.96 | 65.51 | 56.65 | 59.21 |
| PDF | **79.94** | **78.42** | **74.40** | **72.64** | 63.09 | 60.31 |
| DAR + Late Fusion | 78.32 | 76.85 | 74.10 | 72.05 | 63.85 | 61.27 |
| DAR + Concat | 74.90 | 73.41 | 69.47 | 67.59 | **64.14** | **62.12** |

against LNLN, the most recent baseline, the corresponding Acc-2 margin is 2.21 percentage points. The Acc-7 on MOSEI and MAE on MOSI are sub-optimal but very close to the best, and given the inherent variance of corrupted data, DAR achieves the best overall robustness profile. Table 4 extends the conclusion to SIMS (Chinese), where DAR leads on Acc-5, F1, MAE, and Corr. Fig. 6 plots Acc-2 and Acc-5 across the full range of test-time missing rates, showing that DAR's advantage is most pronounced in the moderate-to-high missing regime, precisely where the effective corruption shift relative to training is largest, consistent with Corollary 1.

### 5.3 Robustness under Additive Gaussian Noise

A central claim of this work is that the same DAR pipeline addresses both missing and noisy corruption (Remark 1). We test it by moving to a different corruption operator, a different input structure, and a different instantiation, while leaving the objective of Eq. (40) untouched.

**Results on MVSA.** Table 5 shows that, under increasing noise, traditional static fusion (Late Fusion, Concat) degrades rapidly, uncertainty-aware methods (QMF, PDF) mitigate the drop somewhat, and the DAR-augmented variants provide the strongest robustness in the moderate-to-high noise regimes while remaining competitive on clean data. Critically, both DAR + Late Fusion and DAR + Concat consistently improve over their respective vanilla counter-

TABLE 6
Robustness on **NYU Depth V2** and **SUN RGB-D** under Gaussian noise. The plug-in nature of DAR is most evident in the high-noise regime. Best results are in **bold**, second best are underlined.

| Method | NYU Depth V2 | | | | | | SUN RGB-D | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\sigma = 0$ | | $\sigma = 0.5$ | | $\sigma = 1.0$ | | $\sigma = 0$ | | $\sigma = 0.5$ | | $\sigma = 1.0$ | |
| | Avg | Worst | Avg | Worst | Avg | Worst | Avg | Worst | Avg | Worst | Avg | Worst |
| RGB (uni.) | 63.30 | 62.54 | 53.12 | 50.31 | 45.46 | 42.20 | 56.78 | 56.51 | 48.40 | 47.16 | 42.94 | 41.02 |
| Depth (uni.) | 62.65 | 61.01 | 50.95 | 42.81 | 44.13 | 35.93 | 52.99 | 51.32 | 37.81 | 35.63 | 33.07 | 30.41 |
| Late Fusion | 69.14 | 68.35 | 59.63 | 53.98 | 51.99 | 44.95 | 62.09 | 60.55 | 52.44 | 50.83 | 47.33 | 44.60 |
| Late Fusion with noise(0.5) | 69.69 | 68.81 | 50.42 | 42.81 | 47.07 | 38.83 | 61.46 | 61.02 | 44.98 | 43.19 | 38.62 | 33.01 |
| Concat | 70.30 | 69.42 | 59.97 | 55.89 | 53.20 | 47.71 | 61.90 | 61.19 | 52.69 | 50.61 | 45.64 | 42.95 |
| Concat with noise(0.5) | 71.07 | 70.18 | 50.72 | 47.10 | 48.57 | 43.43 | **62.10** | **61.64** | 44.72 | 42.22 | 40.23 | 38.23 |
| MMTM | 71.04 | 70.18 | 60.37 | 56.73 | 52.28 | 46.18 | 61.72 | 60.94 | 51.86 | 50.80 | 46.03 | 44.28 |
| TMC | 71.06 | 69.57 | 61.04 | 58.72 | 53.36 | 49.23 | 60.68 | 60.31 | 51.24 | 49.45 | 45.66 | 41.60 |
| QMF | 70.09 | 68.81 | 61.62 | 58.87 | 55.60 | 51.07 | 62.09 | 61.30 | 53.40 | 52.07 | 48.58 | 47.50 |
| PDF | 70.04 | 68.90 | 65.72 | 63.91 | 59.44 | 56.37 | 61.30 | 61.14 | **58.53** | **57.62** | 53.95 | 42.31 |
| DAR + Late Fusion | **71.52** | **71.10** | **67.95** | **67.20** | **65.61** | **64.60** | 61.22 | 60.64 | 58.35 | 57.64 | **56.27** | **54.95** |
| DAR + Concat | 70.80 | 70.34 | 66.43 | 64.33 | 63.42 | 60.74 | 59.96 | 59.48 | 57.24 | 56.24 | 51.19 | 48.74 |

parts, demonstrating that the decouple–restore–reconstruct principle transfers even in its purely feedforward form.

**Results on RGB-D scene recognition.** Table 6 extends the conclusion to RGB and depth modalities, which are structurally very different from language and vision. The benefits become *more* pronounced in the high-noise regime ($\sigma = 1.0$), matching Corollary 1: as the pseudo-shared corruption shift $D_p$ grows, equivalently as training-time cross-modal noise correlations become less reliable, the margin over the best baseline widens from $0.45$ to $2.23$ and then $6.17$ points on NYU Depth V2, while baselines remain exposed to the shortcut.

**Discussion: restoration vs. uncertainty weighting.** QMF and PDF address noisy modalities by *down-weighting* unreliable predictions at inference, guided by uncertainty scores. This works best when an entire modality is unreliable. DAR takes a complementary approach: it *reconstructs* a clean estimate using cross-modal redundancy *within* each modality. The two strategies are not mutually exclusive — uncertainty weighting decides *how much* to trust each modality, while DAR determines *what* to trust within each. Combining both is a natural future direction.

TABLE 7
Ablation study on MOSI. $\mathcal{L}_{\text{boot}}$ denotes the task loss applied to the bootstrap branch.

| Variant | Acc-7 | Acc-5 | Acc-2 | F1 | MAE↓ | Corr |
|---|---|---|---|---|---|---|
| w/o $\mathcal{L}_{\text{sim}}$ | 34.14 | 38.42 | 71.50 / 72.71 | 71.30 / 72.62 | 1.084 | 0.505 |
| w/o $\mathcal{L}_{\text{diff}}$ | 34.28 | 38.27 | 71.54 / 72.85 | 71.48 / 72.62 | 1.089 | 0.507 |
| w/o $\mathcal{L}_{\text{sim}}$&$\mathcal{L}_{\text{diff}}$ | 34.15 | 38.35 | 71.32 / 72.46 | 71.10 / 72.35 | 1.113 | 0.504 |
| w/o $\mathcal{L}_{\text{recon}}$ | 33.57 | 38.31 | 71.02 / 72.20 | 70.45 / 71.13 | 1.123 | 0.493 |
| w/o $\mathcal{L}_{\text{boot}}$ | 33.03 | 36.93 | 70.50 / 72.26 | 69.90 / 71.80 | 1.123 | 0.475 |
| **Full DAR** | **34.47** | **38.65** | **71.60 / 73.18** | **71.51 / 73.15** | **1.069** | **0.520** |

TABLE 8
Loss ablation of DAR + Late Fusion under Gaussian noise on NYU Depth V2 (feedforward instantiation).

| Variant | $\sigma = 0$ | | $\sigma = 0.5$ | | $\sigma = 1.0$ | |
|---|---|---|---|---|---|---|
| | Avg | Worst | Avg | Worst | Avg | Worst |
| Late Fusion (baseline) | 69.14 | 68.35 | 59.63 | 53.98 | 51.99 | 44.95 |
| DAR w/o $\mathcal{L}_{\text{sim}}$ | 70.22 | 68.81 | 65.50 | 64.33 | 60.71 | 58.07 |
| DAR w/o $\mathcal{L}_{\text{diff}}$ | 70.37 | 69.27 | 66.24 | 64.33 | 62.04 | 57.87 |
| DAR w/o $\mathcal{L}_{\text{recon}}$ | 70.83 | 70.18 | 67.28 | 66.67 | 64.17 | 62.88 |
| **Full DAR** | **71.52** | **71.10** | **67.95** | **67.20** | **65.61** | **64.60** |

### 5.4 Ablation and Sensitivity Analysis

Having established that DAR helps in both regimes, we now ask which parts of the objective are responsible, and whether the answer is the same in the two regimes. We remove one loss term at a time from $\mathcal{L}_{\text{total}}$ and re-train, first on MOSI under erasure and then on NYU Depth V2 under Gaussian noise, so that the two attributions can be read side by side.

#### 5.4.1 Loss Components under Erasure Corruption

Removing $\mathcal{L}_{\text{sim}}$ or $\mathcal{L}_{\text{diff}}$ weakens the decoupling, allowing common and specific channels to absorb each other's information; in Table 7 performance drops modestly but consistently across metrics, in line with Theorem 1: these two losses implement the constraints that define the restricted DAR hypothesis class. Removing $\mathcal{L}_{\text{recon}}$ eliminates the explicit alignment between restored and clean features and significantly hurts MAE and Corr, confirming that the supervised restoration is essential. The largest drop comes from removing $\mathcal{L}_{\text{boot}}$: without bootstrap supervision, the encoder preferentially produces representations that minimize reconstruction distance even at the cost of task-discriminativeness, consistent with the encoder-degradation issue (limitation L2) discussed in Section 1.

#### 5.4.2 Loss Components under Gaussian Noise

Table 8 repeats the ablation in the feedforward regime. Every variant still improves substantially on the Late Fusion baseline, so none of the removals collapses the model; the ordering, however, is informative. The cross-modal alignment term is the most important here: without $\mathcal{L}_{\text{sim}}$ the average accuracy at $\sigma = 1.0$ falls by $4.90$ points, against $3.57$ for $\mathcal{L}_{\text{diff}}$ and $1.44$ for $\mathcal{L}_{\text{recon}}$. The full model is best at every noise level.

TABLE 9
Sensitivity to the training-time missing rate $k$ on MOSI. The selected setting $k = 0.3$ is shown in **bold**; it is the value used for MOSI in Tables 3 and 7. Best result per column in **bold**.

| Setting | Acc-7 | Acc-5 | Acc-2 | F1 | MAE↓ | Corr |
|---|---|---|---|---|---|---|
| $k = 0.0$ | 31.84 | 35.45 | 68.99 / 71.03 | 66.39 / 63.10 | **1.069** | 0.514 |
| $k = 0.2$ | 33.18 | 37.88 | 70.57 / 71.06 | 70.57 / 70.56 | 1.160 | 0.502 |
| $\boldsymbol{k = 0.3}$ | **34.47** | **38.65** | **71.60 / 73.18** | **71.51 / 73.15** | **1.069** | **0.520** |
| $k = 0.4$ | 32.95 | 36.39 | 71.22 / 72.73 | 70.98 / 72.62 | 1.078 | 0.515 |
| $k = 0.6$ | 30.12 | 32.58 | 70.63 / 71.99 | 70.27 / 71.75 | 1.118 | 0.475 |
| $k = 0.8$ | 24.56 | 24.64 | 69.16 / 70.96 | 67.50 / 69.51 | 1.173 | 0.460 |

#### 5.4.3 What the Two Regimes Share, and Where They Differ

The two ablations agree on the qualitative conclusion — every term contributes, and the full objective is best on every metric at every corruption level — but they disagree on which term matters *most*. Under erasure, the bootstrap branch dominates; under Gaussian noise, the cross-modal alignment term does. The two-channel view of Theorem 2 offers a reading of this reversal. In the sequential regime a partially erased stream still carries usable intra-modal context in its surviving frames, so the specific channel remains informative and the binding constraint is instead keeping the encoder task-discriminative — exactly what $\mathcal{L}_{\text{boot}}$ enforces. In the static regime each modality is a single vector: there is no temporal context to fall back on, the specific channel has little intra-modal structure left to exploit, and the common channel must carry most of the recoverable evidence, which makes cross-modal alignment the critical term. The framework is therefore not merely transferable across the two regimes; its components re-weight themselves in the direction the analysis predicts.

#### 5.4.4 Sensitivity to the Training-Time Missing Rate

The training-time missing rate $k$ is a critical hyperparameter: too small a $k$ degenerates the model into a vanilla multimodal fusion network with no chance to learn restoration; too large a $k$ corrupts both common and specific signals beyond recoverability. We sweep $k \in \{0, 0.2, 0.3, 0.4, 0.6, 0.8\}$ on MOSI and report results in Table 9.

The classification metrics follow the expected single-peaked pattern with the maximum at $k = 0.3$: too small a $k$ underutilizes the restoration capacity, while too large a $k$ destroys the cross-modal redundancy that the common channel relies on. The regression metrics are flatter — MAE at $k = 0$ already matches the best value — so the trend is clearest on the classification metrics, and most visibly on F1, where $k = 0$ trails the peak by $5.12$ and $10.05$ points under the two binarization schemes. The peak is not sharp: every setting in $[0.2, 0.4]$ stays within $1.52$ Acc-7 points of the optimum, so $k$ does not require fine tuning, and only the extreme settings $k \geq 0.6$ degrade the model appreciably. In practice $k$ should match the expected test-time corruption distribution; we selected $k = 0.3$ for MOSI and $k = 0.4$ for MOSEI on validation data. Adaptive missing-rate selection (e.g. modality-specific or curriculum-based) is a promising direction for future work.

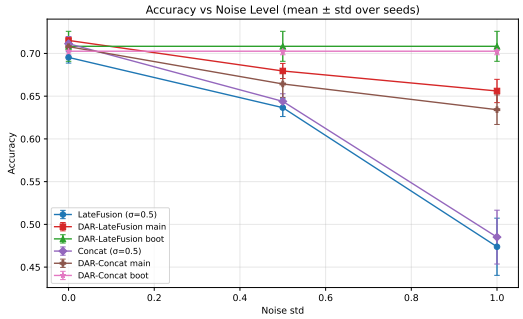



Fig. 7. Test accuracy on NYU Depth V2 as a function of the Gaussian noise standard deviation $\sigma$. Baselines are trained with $\sigma = 0.5$ noise augmentation; "boot" denotes the bootstrap branch of the DAR variants, which reads clean-input features and is therefore nearly independent of $\sigma$.

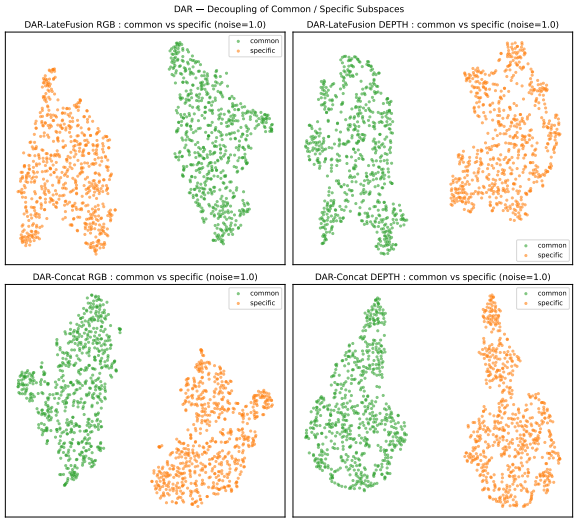

Fig. 8. t-SNE of the fused features on NYU Depth V2 across noise levels $\sigma \in \{0, 0.5, 1.0\}$ (columns) for the four compared methods (rows). Color encodes ground-truth class. DAR variants retain visible cluster structure even at $\sigma = 1.0$.

### 5.5 Representation-Level Analysis

The results so far establish *that* DAR is robust and *which* losses matter. This subsection asks *how*, by inspecting the learned representations directly on NYU Depth V2 across the same noise sweep $\sigma \in \{0, 0.5, 1.0\}$ used in Section 5.3. Throughout, we contrast the two static baselines (Late Fusion, Concat) with their DAR-augmented variants.

**Accuracy degradation curve.** Fig. 7 plots mean accuracy with one-standard-deviation seed bars against $\sigma$. Both static baselines lose more than twenty points between $\sigma = 0$ and $\sigma = 1.0$, whereas the DAR variants lose under eight, so the


(a) Cross-modal alignment in the common subspace.




(b) Decoupling of common vs. specific subspaces.

Fig. 9. Structural diagnostics of the DAR latent space at $\sigma = 1.0$ on NYU Depth V2. (a) Common features of RGB (red) and Depth (blue) form spatially adjacent, class-aligned regions, evidencing the effect of $\mathcal{L}_{\text{sim}}$. (b) Within each modality, common (green) and specific (orange) features remain well separated, the geometric counterpart of $\mathcal{L}_{\text{diff}}$.

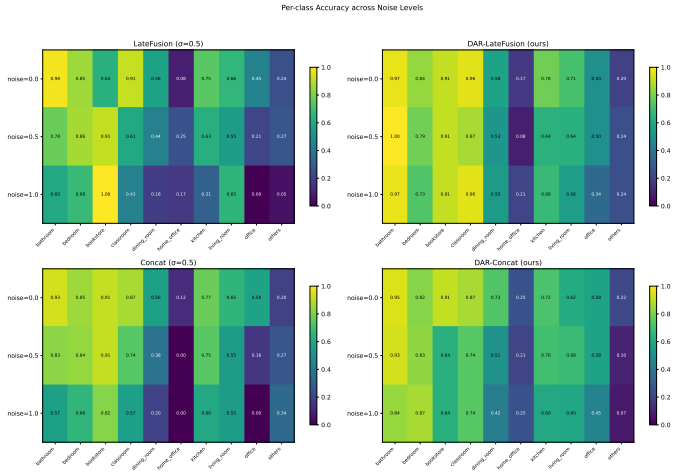



Fig. 10. Per-class accuracy on NYU Depth V2 across noise levels. DAR variants are notably more uniform across the ten scene categories, in particular preserving accuracy on worst-case classes (e.g., *home_office*, *others*) where the baselines collapse to near-zero at $\sigma = 1.0$.

gap at the highest noise level exceeds fifteen accuracy points for DAR-LateFusion. The bootstrap branch ("boot"), which reads the clean-input features, is by construction almost flat in $\sigma$ and thus serves as a noise-free reference: the main branch stays within a few points of it even at $\sigma = 1.0$, indicating that the auxiliary supervision of Section 4 does not come at the expense of main-branch quality.

**Fused-feature t-SNE.** Fig. 8 shows t-SNE projections of the final fused representations colored by ground-truth class. As $\sigma$ grows from 0 to 1.0, the baseline embeddings collapse into a diffuse cloud in which class boundaries are no longer discernible, whereas the DAR variants retain visibly preserved cluster structure even at the highest noise level. This is the representational counterpart of Theorem 2: bounding common–specific leakage limits how strongly the decoded representation can move with the noise-induced shift.

**Cross-modal alignment and decoupling.** Fig. 9 dissects the DAR latent space under heavy noise. Fig. 9a shows that the common embeddings of the two modalities form spatially adjacent and cross-aligned clusters that share class structure — direct evidence that $\mathcal{L}_{\text{sim}}$ drives the two common channels to agree on shared content. Fig. 9b further confirms that, for each modality, the common and specific subspaces remain well separated, which is the geometric counterpart of the MI-based $\mathcal{L}_{\text{diff}}$ objective. Cross-modal alignment and common/specific decoupling are thus *both* achieved even in the high-noise regime, validating the core design hypothesis of DAR and the two-channel mechanism formalized in Theorem 2.

**Per-class robustness.** Fig. 10 displays per-class accuracy heatmaps. DAR variants exhibit markedly more uniform accuracy across the ten scene categories, in particular preserving accuracy on worst-case classes where the baselines collapse at $\sigma = 1.0$. This per-class uniformity directly explains the worst-case improvements reported in Table 6: the gains of DAR are distributed across the label space rather than concentrated on a few easy categories.

**Summary.** These diagnostics jointly corroborate the quantitative findings of Sections 5.3 and 5.4 from a representational standpoint: DAR's decouple–restore–reconstruct pipeline preserves cross-modal alignment, common/specific separation, and class-level discriminative structure under heavy noise, providing the geometric mechanism that underlies its robustness gain.

## 6 Conclusion

We have introduced **Decoupled-Adaptive Reconstruction (DAR)**, a unified framework for multimodal learning under low-quality data that subsumes both missing-modality and

noisy-modality regimes. The framework rests on a simple decouple–restore–reconstruct principle: each modality's representation is factorized into a modal-common and a modal-specific component using mutual-information-based contrastive objectives; the two components are restored from the corrupted input by structurally distinct heads exploiting cross-modal redundancy and intra-modal consistency respectively; and a gated fusion adaptively reconstructs the clean signal.

**Theoretically**, we derived a generalization analysis for the setting in which corruption statistics shift between training and test: the mutual-information constraints define a restricted hypothesis class that never increases statistical complexity (Theorem 1); under a balanced two-channel coupling condition, the common–specific constraint bounds the sensitivity of the learned representation to environment-dependent pseudo-shared shortcuts (Theorem 2); and combining the fitting, estimation, and shift-sensitivity terms yields a transparent sufficient condition for a strict test-risk advantage over corruption-augmented training (Theorem 3), covering missing and noise corruption through a single corruption operator.

**Empirically**, on three multimodal sentiment datasets under varying missing rates (MOSI, MOSEI, SIMS) and on three image–text and RGB-D datasets under Gaussian noise corruption (MVSA, NYU Depth V2, SUN RGB-D), DAR consistently outperforms strong static-fusion, dynamic-fusion, and reconstruction-based baselines. Notably, these results are obtained with two structurally different instantiations of the same role-defined framework — an attention/recurrent realization for sequential inputs and a purely feedforward realization for static feature vectors — confirming that the gains follow from the decouple–restore–reconstruct principle and its mutual-information constraints rather than from any particular architectural choice. The advantage grows with corruption intensity, consistent with the increasing-shift regime described in Corollary 1.

**Future directions.** Promising extensions include adaptive or curriculum-based corruption schedules that mirror the expected test distribution; combining DAR's within-modality restoration with uncertainty-aware cross-modality reweighting paradigms such as QMF and PDF; and relaxing the fixed-corruption-distribution assumption of our analysis to broader covariate shifts on real-world domain-shifted benchmarks.

## REFERENCES


[1] T. Liang, G. Lin, L. Feng, Y. Zhang, and F. Lv, "Attention is not enough: Mitigating the distribution discrepancy in asynchronous multimodal sequence fusion," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8128–8136, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244074135

[2] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2554–2562.

[3] Y. Zhang, Y. Liu, S. Wu, K. Zhang, X. Liu, Q. Liu, and E. Chen, "Leveraging entity information for cross-modality correlation learning: The entity-guided multimodal summarization," in *Findings of the Association for Computational Linguistics ACL 2024*, 2024, pp. 9851–9862.

[4] Y. Liu, K. Zhang, Z. Huang, K. Wang, Y. Zhang, Q. Liu, and E. Chen, "Enhancing hierarchical text classification through knowledge graph integration," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 5797–5810.

[5] Z. Yuan, W. Li, H. Xu, and W. Yu, "Transformer-based feature reconstruction network for robust multimodal sentiment analysis," in *Proceedings of the 29th ACM International Conference on Multimedia*, ser. MM '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 4400–4407. [Online]. Available: https://doi.org/10.1145/3474085.3475585

[6] H. Zhang, W. Wang, and T. Yu, "Towards robust multimodal sentiment analysis with incomplete data," in *The Thirty-eighth Annual Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

[7] Z. Lian, L. Chen, L. Sun, B. Liu, and J. Tao, "GCNet: Graph completion network for incomplete multimodal learning in conversation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 7, pp. 8419–8432, 2023.

[8] Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng, "Provable dynamic fusion for low-quality multimodal data," in *Proceedings of the International Conference on Machine Learning*, 2023, pp. 41 753–41 769.

[9] B. Cao, Y. Xia, Y. Ding, C. Zhang, and Q. Hu, "Predictive dynamic fusion," in *Proceedings of the International Conference on Machine Learning*, 2024, pp. 5608–5628.

[10] T. Mittal, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 02, 2020, pp. 1359–1367.

[11] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *International Conference on Learning Representations*, 2019.

[12] J. Zhao, R. Li, and Q. 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*, 2021, pp. 2608–2618.

[13] Z. Yuan, Y. Liu, H. Xu, and K. Gao, "Noise imitation based adversarial training for robust multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 26, pp. 529–539, 2024.

[14] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *Proceedings of the International Conference on Learning Representations*, 2021.

[15] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, and L. Carin, "Club: A contrastive log-ratio upper bound of mutual information," in *International conference on machine learning*. PMLR, 2020, pp. 1779–1788.

[16] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, M. Palmer, R. Hwa, and S. Riedel, Eds. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 1103–1114. [Online]. Available: https://aclanthology.org/D17-1115/

[17] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[18] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[19] S. Mai, S. Xing, and H. Hu, "Locally confined modality fusion network with a global perspective for multimodal human affective computing," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 122–137, 2020.

[20] W. Rahman, M. K. Hasan, S. Lee, A. Bagher Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 2359–2369. [Online]. Available: https://aclanthology.org/2020.acl-main.214/

[21] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 10 790–10 797.

[22] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, pp. 3371–3408, 2010.

[23] M. Ma, J. Ren, L. Zhao, S. Tulyakov, C. Wu, and X. Peng, "SMIL: Multimodal learning with severely missing modality," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 3, 2021, pp. 2302–2310.

[24] M. Li, D. Yang, Y. Lei, S. Wang, S. Wang, L. Su, K. Yang, Y. Wang, M. Sun, and L. Zhang, "A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities," in *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2025. [Online]. Available: https://doi.org/10.1609/aaai.v38i9.28871

[25] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 2, pp. 742–754, 2025.

[26] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[27] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2402–2415, 2022.

[28] Y. Lin, Y. Gou, X. Liu, J. Bai, J. Lv, and X. Peng, "Dual contrastive prediction for incomplete multi-view representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4447–4461, 2023.

[29] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification with dynamic evidential fusion," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 2551–2566, 2023.

[30] Z. Xue and R. Marculescu, "Dynamic multimodal fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2023, pp. 2575–2584.

[31] X. Wang, H. Chen, S. Tang, Z. Wu, and W. Zhu, "Disentangled representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 9677–9696, 2024.

[32] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," in *International Conference on Learning Representations*, 2019.

[33] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and -specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1122–1131. [Online]. Available: https://doi.org/10.1145/3394171.3413678

[34] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 6631–6640.

[35] J. Yang, Y. Yu, D. Niu, W. Guo, and Y. Xu, "ConFEDE: Contrastive feature decomposition for multimodal sentiment analysis," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 7617–7630. [Online]. Available: https://aclanthology.org/2023.acl-long.421/

[36] Y. Xia, H. Huang, J. Zhu, and Z. Zhao, "Achieving cross modal generalization with multimodal unified representation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[37] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. Carneiro, "Multi-modal learning with missing modality via shared-specific feature modelling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 15 878–15 887.

[38] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[39] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *Journal of Machine Learning Research*, vol. 3, pp. 463–482, 2002.

[40] M. Yang, K. Zhang, Y. Ye, Y. Zhang, R. Yu, and M. Hou, "Decoupling and reconstructing: A multimodal sentiment analysis framework towards robustness." in *IJCAI*, 2025, pp. 6803–6811.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[42] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[43] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[44] A. Bagher Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, I. Gurevych and Y. Miyao, Eds. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2236–2246. [Online]. Available: https://aclanthology.org/P18-1208/

[45] W. Yu, H. Xu, F. Meng, Y. Zhu, Y. Ma, J. Wu, J. Zou, and K. Yang, "CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3718–3727.

[46] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," in *Proceedings of the International Conference on Multimedia Modeling*, 2016, pp. 15–27.

[47] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proceedings of the European Conference on Computer Vision*, 2012, pp. 746–760.

[48] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 567–576.

[49] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. 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*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 756–767. [Online]. Available: https://aclanthology.org/2023.emnlp-main.49/

[50] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2019, pp. 4171–4186.

[51] B. McFee, C. Raffel, D. Liang, D. P. W. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *SciPy*, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:33504

[52] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018, pp. 59–66.

[53] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 9180–9192. [Online]. Available: https://aclanthology.org/2021.emnlp-main.723/

[54] D. Wang, S. Liu, Q. Wang, Y. Tian, L. He, and X. Gao, "Cross-modal enhancement network for multimodal sentiment analysis," *IEEE Transactions on Multimedia*, vol. 25, pp. 4909–4921, 2022.

[55] D. Wang, X. Guo, Y. Tian, J. Liu, L. He, and X. Luo, "TETFN: A text enhanced transformer fusion network for multimodal sentiment analysis," *Pattern Recognition*, vol. 136, p. 109259, 2023.

[56] H. Mao, Z. Yuan, H. Xu, W. Yu, Y. Liu, and K. Gao, "M-SENA: An integrated platform for multimodal sentiment analysis," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, V. Basile, Z. Kozareva, and S. Stajner, Eds. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 204–213. [Online]. Available: https://aclanthology.org/2022.acl-demo.20/

[57] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for CNN fusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 289–13 299.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[59] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

**Kai Zhang** (Member, IEEE) is an Associate Researcher at the University of Science and Technology of China (USTC). He received his Ph.D. degree in Computer Science from USTC. His general research areas lie in natural language processing and knowledge discovery. He has published extensively in leading refereed journals and conference proceedings, including IEEE TKDE, ACL, IJCAI, AAAI, KDD, SIGIR, WSDM, and WWW. He regularly serves on the program committees of several top academic conferences and is an active reviewer for prominent journals in his field. He is a member of ACM, SIGIR, AAAI, and CCF. As the first author, he received the CCML 2019 Best Student Paper Award.

**Mingzheng Yang** received his B.Ec. degree in Software Engineering from Beihang University, Beijing, China, in 2023. He is currently working toward the Ph.D degree in the School of Computer Science from University of Science and Technology of China, Hefei, China. His research interests focus on multimodal learning and efficient inference for Large Language Models.

**Qi Liu** (Member, IEEE) received the PhD degree from the University of Science and Technology of China (USTC), Hefei, China, in 2013. He is currently a professor with the School of Artificial Intelligence and Data Science, USTC. His general area of research is data mining and knowledge discovery. He has published prolifically in refereed journals and conference proceedings (e.g., IEEE Transactions on Knowledge and Data Engineering, ACM Transactions on Information Systems, KDD). He is an associate editor of IEEE Transactions on Big Data and Neurocomputing. He was the recipient of the KDD'18 Best Student Paper Award and ICDM'11 Best Research Paper Award. He is a member of the Alibaba DAMO Academy Young Fellow program. He was also the recipient of the Outstanding Youth Science Foundation of China in 2019.

**Enhong Chen** (Fellow, IEEE) received the PhD degree from the University of Science and Technology of China (USTC), Hefei, China, in 1996. He is currently a professor and the vice dean with the School of Computer Science, USTC. He has published more than 200 papers in refereed conferences and journals, including IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Mobile Computing, ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), IEEE International Conference on Data Mining (ICDM), Conference on Neural Information Processing Systems, and ACM International Conference on Information and Knowledge Management. His current research interests include data mining and machine learning, social network analysis, and recommender systems. He was a recipient of the Best Application Paper Award at KDD 2008, the Best Research Paper Award at ICDM 2011, and the Best of SIAM International Conference on Data Mining (SDM)-2015. He has served on the program committees of numerous conferences, including KDD, ICDM, and SDM.