arXiv:2602.16197v2 [cs.LG] 9 Mar 2026

MODALIMMUNE: IMMUNITY DRIVEN UNLEARNING VIA SELF DESTRUCTIVE TRAINING

Rong Fu*
University of Macau
mc46603@um.edu.mo

Jia Yee Tan
Renmin University of China
tanjiayi2002@ruc.edu.cn

Zijian Zhang
University of Pennsylvania
zzjharry@alumni.upenn.edu

Ziming Wang
Zhejiang University
w986827512@zuaa.zju.edu.cn

Zhaolu Kang
Peking University
kangzl9966@gmail.com

Muge Qi
Peking University
2301210659@stu.pku.edu.cn

Shuning Zhang
Tsinghua University
zsn23@mails.tsinghua.edu.cn

Simon Fong
University of Macau
ccfong@um.edu.mo

March 10, 2026

ABSTRACT

Multimodal systems are vulnerable to partial or complete loss of input channels at deployment, which undermines reliability in real-world settings. This paper presents ModalImmune, a training framework that enforces modality immunity by intentionally and controllably collapsing selected modality information during training so the model learns joint representations that are robust to destructive modality influence. The framework combines a spectrum-adaptive collapse regularizer, an information-gain guided controller for targeted interventions, curvature-aware gradient masking to stabilize destructive updates, and a certified Neumann-truncated hyper-gradient procedure for automatic meta-parameter adaptation. Empirical evaluation on standard multimodal benchmarks demonstrates that ModalImmune improves resilience to modality removal and corruption while retaining convergence stability and reconstruction capacity.

Keywords multimodal robustness, immunity-driven unlearning, self-destructive training, spectral collapse, hyper-gradient adaptation

1 Introduction

Multimodal learning integrates complementary signals from heterogeneous channels such as text, audio and vision to achieve higher prediction accuracy and richer inference. Despite advances, models that assume complete and well-aligned inputs often fail in realistic deployments where one or more modalities may be missing, corrupted, or unavailable due to sensor faults, privacy constraints, communication dropouts, or adversarial manipulation. This fragility hinders adoption of multimodal systems in safety-critical and resource-constrained scenarios and motivates methods that improve robustness to missing or destructive modalities [1, 2, 3].

Existing remedies follow several directions. Generative imputation reconstructs missing channels from available inputs using conditional generators or diffusion models, which can restore performance in some cases but risks hallucination and incurs additional computational cost [2, 1]. Architectural strategies build modular fusion, expert routing or adaptive gating to reduce over-reliance on any single modality, yet many such solutions are trained for specific missing patterns or require extra components that constrain generality [4, 3]. Information-centric objectives and contrastive schemes aim

*Corresponding author: mc46603@um.edu.mo

to shape modality-invariant embeddings and to suppress spurious unimodal cues, but they do not explicitly enforce immunity against a modality that may become destructive when present [5, 6, 7].

Three limitations persist. First, imputational approaches may introduce low-fidelity signals and impose runtime burdens that limit practicality. Second, many adaptive fusion or reconstruction techniques assume fixed missing patterns or add modules that hinder architectural flexibility. Third, there is a lack of principled training protocols that deliberately expose models to controlled destructive interventions so that joint representations become inherently resistant to a modality's harmful influence while preserving task-relevant information from remaining channels.

To address these gaps we introduce ModalImmune, a unified training protocol that operationalizes immunity-driven unlearning via Self Destructive Training. The protocol repeatedly and selectively enforces a controlled collapse on a chosen modality during training so that the encoder and fusion pathways learn to down-weight or ignore destructive information. The collapse is spectrum-adaptive to remove informative directions without destabilizing optimization. An information-gain guided controller prioritizes high-impact interventions to improve training efficiency. Curvature-aware gradient masking prevents unstable gradient ascent when destructive updates are applied. Meta-parameters that govern collapse strength and stabilization are adapted by a certified Neumann-truncated hyper-gradient procedure that provides stable bi-level optimization.

Our contributions are as follows. First, we propose Self Destructive Training as a new paradigm for enforcing modality immunity by targeted, controlled information collapse during model training. Second, we design a spectrum-adaptive collapse regularizer together with an information-gain driven controller that identifies and prioritizes high-impact modality interventions. Third, we develop curvature-aware gradient masking to stabilize destructive updates and a certified Neumann-truncated hyper-gradient algorithm to adapt meta-parameters automatically. Fourth, we provide comprehensive experiments showing that ModalImmune yields representations that maintain predictive performance under modality removal and corruption while preserving convergence behavior and reconstruction capability.

2 Related Work

2.1 Learning with missing modalities

Robust handling of incomplete modality sets is a central concern for contemporary multimodal learning. A broad family of approaches focuses on reconstructing absent channels or learning modality-agnostic representations so that downstream predictors remain effective when some inputs are unavailable. Recent works propose distillation and translation schemes that reconstruct missing semantics from available modalities and align representations across modalities to reduce the performance gap caused by missing data [8, 9, 10]. The correlation-decoupled knowledge distillation framework and hierarchical representation learning both demonstrate that transferring structured cross-sample and category-level knowledge improves robustness under uncertain missingness.

In parallel, medical-imaging and healthcare applications highlight practical constraints where modality absence is common. Masking-based reconstruction ideas that treat missing inputs as masked features and learn to reconstruct them via feature-to-feature mapping have been shown effective for clinical segmentation tasks [11]. These methods demonstrate the value of exposing models during training to realistic missing patterns so they generalize to heterogeneous deployment scenarios.

2.2 Generative imputation and joint latent models

Generative latent-variable models and variational autoencoder families have been widely applied to multimodal imputation and embedding. Joint VAE formulations offer a principled route to infer missing channels while preserving modality-specific structure, and they have been validated on complex biological and sensing datasets [12]. Related lines of work exploit conditional generative models and cross-modal synthesis to produce surrogate inputs that feed into fusion modules; such generative imputation often yields better downstream performance than naive zero- or mean-filling [13, 14].

2.3 Contrastive learning and representation regularization

Contrastive objectives and prototype-guided alignment strategies are now standard tools to obtain discriminative multi-modal embeddings. Methods that combine sample-level contrastive signals with prototype or curriculum mechanisms improve class separation and resilience in low-data settings [15, 16]. Distillation that preserves inter-sample correlations has also been used to reconstruct missing semantics and to regularize student representations for robustness [8]. These results point to a recurring design pattern: learning objectives that explicitly shape inter- and intra-modal geometry confer robustness under partial observability.

2.4 Architectural strategies for partial observability

Architectures tailored for partial modality availability include modular fusion, expert routing and reconstruction-augmented fusion. Graph and hypergraph constructions capture higher-order inter-modal relations and can mitigate the effects of modality noise and absence [17, 11]. Transformer-based translation modules that map non-text modalities into richer textual spaces also yield robust joint features by leveraging pretrained textual priors [18, 7]. Empirical comparisons underscore that combining structured fusion with learned imputation produces more graceful degradation than monolithic fusion designs.

2.5 Information-theoretic and bottleneck regularizers

Imposing information constraints on representations is an effective route to discard nuisance variability while preserving task-relevant signals. The multimodal information bottleneck and related variational formulations reduce redundancy and suppress noisy unimodal components, thereby improving generalization and robustness [16, 6]. These principles are complementary to generative imputation and can be integrated as regularizers to encourage compact, task-focused embeddings.

2.6 Representation collapse and mitigation

Recent analyses have examined the phenomenon of modality collapse, where fusion networks disproportionately rely on a subset of channels and ignore others. Theoretical and empirical studies show that collapse can arise from entanglement in the fusion head or rank bottlenecks in student encoders, and that targeted basis reallocation or cross-modal regularization can alleviate the problem [19]. These insights motivate design choices that explicitly preserve representational capacity per modality and that actively redistribute information when collapse tendencies are detected.

2.7 Optimization, bilevel adaptation and gradient control

Hyperparameter adaptation through bilevel optimization and approximate hyper-gradient estimation has been used to tune meta-parameters that control reconstruction strength and regularization intensity [20, 21, 22]. Practical algorithms exploit truncated Neumann-series approximations to stabilize hyper-gradient computation while keeping compute overhead manageable. Complementary work on gradient masking and routing demonstrates how localized gradient control can limit destructive interference across modules and enable targeted feedback mechanisms [23, 24].

2.8 Bandit selection, information gain and robust evaluation

Adaptive selection of interventions and targeted perturbations benefits from information-aware bandit strategies and causal regret analyses. Information-gain driven arm selection offers a principled mechanism to identify which modality to perturb or mask during training so that the model learns robust conditional dependencies [25, 26, 27]. For evaluation, recent benchmarks and corruption studies systematically probe model robustness under varied missingness and noise regimes, proposing standardized metrics and stress tests to quantify practical resilience [28, 29].

2.9 Positioning of this work

This study introduces a unified training strategy that integrates spectrum-adaptive collapse, curvature-aware gradient regulation, and certified hyper-gradient tuning to achieve modality immunity. Unlike imputation or single-pass distillation, the proposed protocol deliberately applies controlled destructive interventions while retaining reconstruction capability. Compared to Random Modality Dropout, which injects passive noise, our approach actively removes high-impact spectral components, creating a calibrated information deficit. Unlike adversarial training that optimizes worst-case perturbations under norm constraints, our method employs an information-theoretic controller to prioritize impactful collapses and stabilizes updates through curvature gating. These design choices distinguish the framework from existing robustness schemes and establish a principled pathway toward adaptive multimodal resilience.

3 Methodology

We introduce **ModalImmune**, a unified training protocol that promotes modality immunity by deliberately and controllably destroying task-relevant information in a selected modality during training. The protocol operates by alternating standard reconstruction-driven updates with targeted *Self-Destructive Learning* (SDL) phases that enforce spectrum-adaptive collapse on one modality at a time, guided by an information-gain controller and stabilized with

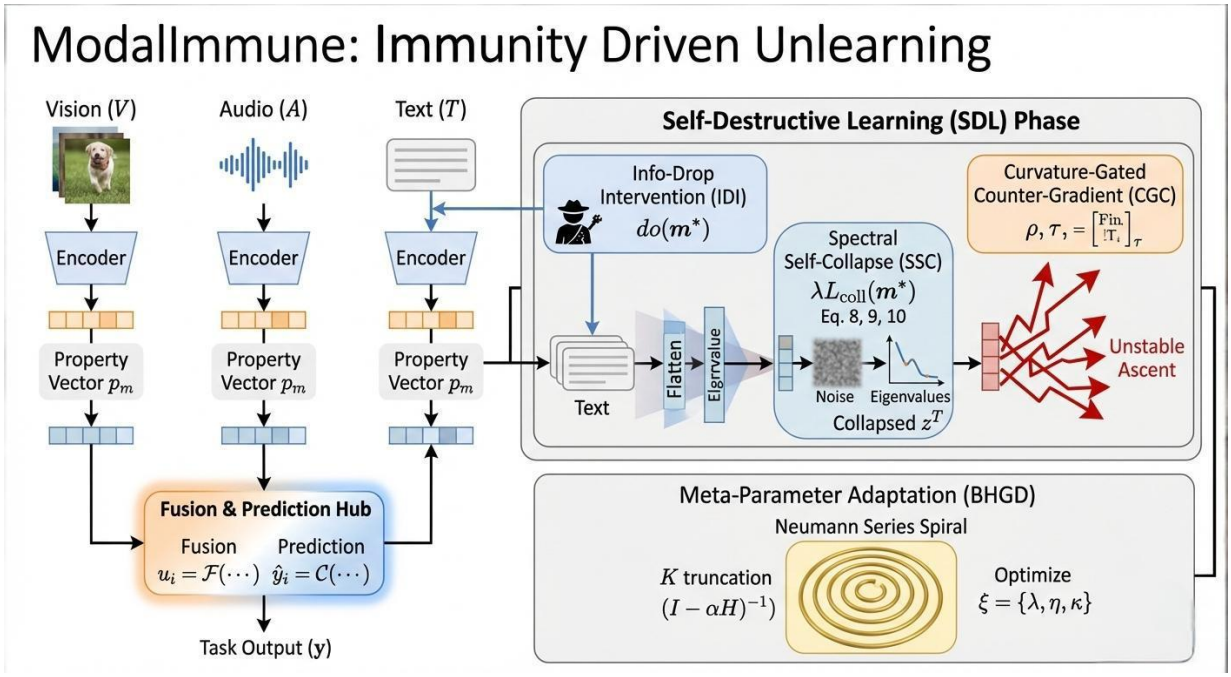


Figure 1: Overview of the **ModalImmune** framework, which treats modality destruction as an active **causal intervention**. The training strategy alternates between standard reconstruction and **Self-Destructive Learning (SDL)** built on three key components: **Info-Drop Intervention (IDI)**, where an EXP3.P bandit controller leverages an information-gain surrogate ℓ_m to adaptively select the target modality $m^\star$; **Spectral Self-Collapse (SSC)**, enforcing an irreversible directional loss through a spectrum-adaptive regularizer L_{coll} combined with a stable-rank penalty; **Curvature-Gated Counter-Gradient (CGC)**, which inspects empirical Fisher information to apply a masking multiplier ρ that mitigates destabilizing gradient ascent. Meanwhile, meta-parameters $\xi = \{\lambda, \eta, \kappa\}$ are optimized via **Bi-level Hyper-Gradient Descent (BHGD)** using a certified **Neumann-truncated** estimator. This unified protocol ensures the **Fusion Hub** and **Task Output** (C) remain robust under counterfactual interventions and adversarial modality collapse.

curvature-aware gradient gating. Meta-parameters that regulate collapse strength and stabilization are adapted online via a certified Neumann-truncated hyper-gradient estimator.

3.1 Self-Destructive Learning as a causal intervention

We view missing or destroyed modalities as active causal interventions rather than passive noise. Let the complete set of modality embeddings be $\{z_m\}_{m\in\mathcal{M}}$ and suppose the label y is generated by a (possibly unknown) structural mapping

$$y \leftarrow g\big(\{z_m\}_{m\in\mathcal{M}}\big), \tag{1}$$

where $g(\cdot)$ denotes the label-generating mechanism. When we intervene on a modality $m^\star$ we apply a do-style operation that replaces $z_{m^\star}$ with a destroyed (collapsed) counterpart, producing the counterfactual

$$\tilde{y} \leftarrow g\big(\{z_m\}_{m\in\mathcal{M}\setminus\{m^\star\}},\ \text{Collapse}(z_{m^\star})\big). \tag{2}$$

where $\text{Collapse}(\cdot)$ denotes a controlled information-destruction operator (defined below). Training under such interventions forces the fusion and prediction pathways to learn representations that are robust to the presence of damaged or adversarial modalities.

3.2 Three design primitives

To operationalize the causal intervention above, ModalImmune relies on three low-level design primitives that are used throughout the subsequent equations and the training algorithm.

Info-Drop Intervention (IDI). An adaptive selector chooses the modality to intervene on in each SDL batch by estimating per-modality information gain and using a bandit controller to prioritize high-impact choices. This avoids wasting destructive steps on low-value targets and concentrates training on modalities that most affect task performance.

Spectral Self-Collapse (SSC). When a modality is chosen, its batch embedding matrix undergoes a controlled spectral collapse that removes dominant directions while preserving global scale. The collapse is stochastic and spectrum-adaptive so that it produces effectively irreversible loss of directional information rather than small perturbations.

Curvature-Gated Counter-Gradient (CGC). To prevent destructive updates from destabilizing optimization, a curvature gate inspects a Gauss–Newton or empirical Fisher approximation and either freezes the encoder gradients for the attacked modality or applies a bounded negative-feedback mask that prevents uncontrolled ascent.

3.3 Notation

Let $\mathcal{M}$ be the set of modalities and $\mathcal{D} = \{(x_i^m)_{m\in\mathcal{M}}, y_i\}_{i=1}^N$ the labeled dataset of N examples. Each modality m has an encoder $f_m(\cdot;\theta_m)$ producing embeddings $z_i^m \in \mathbb{R}^d$. Fusion is performed by $u_i = F(\{z_i^m\}_{m\in\mathcal{S}_i};\phi)$ where $\mathcal{S}_i \subseteq \mathcal{M}$ denotes the modalities available (or synthesized) for sample i, and the task prediction is $\hat{y}_i = C(u_i;\psi)$. The supervised per-sample loss is denoted L_{TASK}, the reconstruction loss by L_{re}, the property-vector loss by L_{pe}, and the contrastive alignment loss by L_{con}.

3.4 Model components and composite objective

Each modality has a learnable property vector $p_m \in \mathbb{R}^d$ that captures sample-invariant traits. A conditional generator $\mathcal{G}_m(\cdot;\vartheta_m)$ can synthesize a surrogate embedding $\hat{z}_i^m$ for a missing modality by combining available embeddings with p_m. A decomposition module optionally maps z_i^m to (s_i^m, μ_i^m) where s_i^m is sample-specific variation and μ_i^m captures invariant characteristics. Back-translation networks $\mathcal{B}_m(\cdot;\beta_m)$ are used together with contrastive alignment to preserve modality-specific signals when needed.

SDL is implemented as a plug-in regularizer. The total training objective for a batch is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{orig}} + \mathbb{I}_{\text{SDL}}\ \lambda\ L_{\text{coll}}(m^\star), \tag{3}$$

where $\mathcal{L}_{\text{orig}}$ aggregates the reconstruction and contrastive terms, $\mathbb{I}_{\text{SDL}} \in \{0,1\}$ indicates whether the current batch is an SDL batch, $\lambda > 0$ weighs the collapse regularizer, and $L_{\text{coll}}(m^\star)$ is the spectral collapse penalty applied to the selected modality $m^\star$.

3.5 Property Vector: Sample-Invariant Modality Signature

The property vector p_m is designed as a modality-specific signature shared across all samples, capturing global characteristics rather than per-utterance variations. Examples include the average spectral slope for acoustic streams or typical color distribution bias for visual encoders. During SDL, when the embedding of the targeted modality is collapsed, the conditional generator G_m combines intact embeddings with this static signature to synthesize a plausible substitute. This ensures that the fusion hub receives inputs aligned with the original semantic scale. In essence, p_m serves as a modality-level prior, whereas the conventional feature z_i^m represents instantaneous, sample-dependent content.

3.6 Training objectives

Training alternates between standard reconstruction-driven updates and SDL updates. For a batch B of size n the reconstruction-driven objective reads

$$\mathcal{L}_{\text{rec}} = \frac{1}{n}\sum_{i\in B} L_{\text{TASK}}\big(C(F(\{z_i^m\}_{m\in\mathcal{S}_i};\phi);\psi), y_i\big) \tag{4}$$

$$+ \alpha L_{\text{re}} + \beta L_{\text{pe}} + \gamma L_{\text{con}}, \tag{5}$$

where $\alpha, \beta, \gamma \geq 0$ are scalar weights. Here L_{TASK} is cross-entropy for classification or mean-squared error for regression.

When SDL is active and modality $m^\star$ is selected, the SDL objective becomes

$$\mathcal{L}_{\text{SDL}} = \frac{1}{n}\sum_{i\in B} L_{\text{TASK}}\big(C(F(\{z_i^m\}_{m\neq m^\star};\phi);\psi), y_i\big) \tag{6}$$

$$+ \gamma L_{\text{con}} + \lambda L_{\text{coll}}(m^\star), \tag{7}$$

which omits $m^\star$ from fusion and enforces the collapse penalty. The per-batch loss used for updates is the interpolation

$$\mathcal{L} = (1 - \mathbb{I}_{\text{SDL}})\,\mathcal{L}_{\text{rec}} + \mathbb{I}_{\text{SDL}}\,\mathcal{L}_{\text{SDL}}. \tag{8}$$

In all formulas $\mathbb{I}_{\text{SDL}}$ denotes the SDL-mode indicator for the batch.

3.7 Spectrum-adaptive collapse regularizer

Let Cov_B be the empirical covariance of the selected-modality embeddings $\{z_i^{m^\star}\}_{i\in B}$ in the current batch. Define the per-batch spectral radius

$$\rho_B = \mathrm{spectral_radius}\big(\{z_i^{m^\star}\}_{i\in B}\big), \tag{9}$$

where the spectral radius returns the largest singular value of the batch matrix. We set a noise scale proportional to embedding magnitude:

$$\sigma_B = \frac{\rho_B}{\sqrt{d}}, \tag{10}$$

where d is the embedding dimension. To obtain a rank-robust penalty when $n < d$ we use the stable-rank

$$\mathrm{stable_rank}(\mathrm{Cov}_B) = \frac{\|\mathrm{Cov}_B\|_*}{\|\mathrm{Cov}_B\|_2}, \tag{11}$$

where $\|\cdot\|_*$ is the nuclear norm and $\|\cdot\|_2$ the operator norm. The collapse regularizer is

$$L_{\mathrm{coll}}(m^\star) = \frac{1}{n}\sum_{i\in B}\big\|z_i^{m^\star} - \varepsilon_i\big\|_2^2 + \eta\cdot\mathrm{stable_rank}(\mathrm{Cov}_B), \tag{12}$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma_B^2 I_d)$ and $\eta \geq 0$ controls the stable-rank penalty. The stochastic perturbation plus stable-rank term yields an effectively irreversible reduction of directional information while keeping overall scale compatible with remaining modalities.

3.8 Curvature-gated gradient masking

Let $\mathcal{F}_{m^\star}$ denote the empirical Fisher (or Gauss–Newton) approximation for encoder $f_{m^\star}$ on the current batch and let $\lambda_{\min}$ be its smallest eigenvalue. Define a curvature threshold

$$\tau = \frac{0.01}{\mathrm{lr}} \tag{13}$$

where lr is the encoder learning rate. The masking multiplier ρ is chosen by

$$\rho = \begin{cases} 0, & \lambda_{\min} < -\tau, \\ -\kappa, & \lambda_{\min} \geq -\tau, \end{cases} \tag{14}$$

where $\kappa > 0$ is the negative-feedback scale. When $\rho = 0$ gradients for the attacked encoder are frozen; when $\rho = -\kappa$ a controlled negative feedback is applied to the encoder updates to avoid ascent-based divergence.

3.9 Hyper-gradient estimation via certified Neumann truncation

We treat meta-parameters $\xi = \{\lambda, \eta, \kappa\}$ as latent variables and approximate hyper-gradients of the validation loss $\mathcal{L}_{\mathrm{val}}(\xi)$ by

$$\nabla_\xi \mathcal{L}_{\mathrm{val}} \approx \frac{\partial \mathcal{L}_{\mathrm{val}}}{\partial \theta}\,(I - \alpha H)^{-1}\,\frac{\partial^2 \mathcal{L}_{\mathrm{train}}}{\partial \theta \partial \xi}\bigg|_{\theta^*(\xi)}, \tag{15}$$

where $H = \partial^2 \mathcal{L}_{\mathrm{train}}/\partial\theta^2$ is the Hessian approximation and α is the inner learning rate. The inverse $(I - \alpha H)^{-1}$ is approximated by a Neumann series truncated at depth K, and K is chosen online via a doubling rule until the truncation residual is below a prescribed tolerance relative to stochastic estimation noise. This yields a verifiable truncation certificate and stabilizes BHGD updates.

3.10 Information-gain surrogate for bandit selection

To supply a causal-relevant reward to the EXP3.P bandit we approximate per-modality information gain on batch B with the surrogate

$$\ell_m \approx \frac{1}{|B|}\sum_{i\in B}\Big(L_{\mathrm{task}}\big(\mathcal{M}\big) - L_{\mathrm{task}}\big(\mathcal{M}\setminus\{m\}\big)\Big)^2, \tag{16}$$

where $L_{\mathrm{task}}(\mathcal{S})$ denotes the task loss computed when using modalities in the set $\mathcal{S}$. This proxy is dimensionless and approximates the conditional information $\mathcal{I}(z^m; y \mid z^{\mathcal{M}\setminus m})$ using current model predictions.

Algorithm 1: ModalImmune: INFO-GAIN bandit + BHGD

Input: dataset $\mathcal{D}$, modalities $\mathcal{M}$, SDL probability p_{sdl}, weights α, β, γ, bandit and BHGD hyper-params

Initialize EXP3.P weights for arms $m \in \mathcal{M}$ and meta-parameters ξ;
for *each epoch* **do**
  for *each mini-batch* $B \subset \mathcal{D}$ **do**
    sample $u \sim \text{Uniform}(0, 1)$;
    if $u < p_{\text{sdl}}$ **then**
      set $\mathbb{I}_{\text{SDL}} \leftarrow 1$;
      select arm $m^\star$ via EXP3.P using recent ℓ_m estimates;
      set $\mathcal{S}_i \leftarrow \mathcal{M} \setminus \{m^\star\}$ for all $i \in B$;
      forward-pass using modalities in $\mathcal{S}_i$, compute $L_{\text{coll}}(m^\star)$ by Eq. (12);
      compute gradients, apply curvature-gated masking and step the optimizer;
      observe $\ell_{m^\star}$ and update EXP3.P;
    else
      set $\mathbb{I}_{\text{SDL}} \leftarrow 0$;
      perform standard reconstruction-driven forward-pass (including generator synthesis for randomly masked modalities);
      update parameters by minimizing $\mathcal{L}_{\text{rec}}$;
    periodically compute approximate hyper-gradient and update ξ via BHGD;

3.11 Training algorithm (summary)

3.12 Compatibility certificate and Lipschitz proxy

To ease encoder replacement we provide a conservative spectral upper bound on the fusion Lipschitz constant for modality m:

$$L_F^{\langle m \rangle} \;=\; \sup_{x^m} \left\| \nabla_{x^m} F \circ f_{\theta_m}(x^m) \right\|_2 \le \|\mathbf{W}_F\|_2 \cdot \|\mathbf{W}_{\theta_m}\|_2, \tag{17}$$

where $\mathbf{W}_F$ is the first linear layer (in the fusion network) receiving modality m and $\mathbf{W}_{\theta_m}$ is the last linear layer of encoder f_{θ_m}. Both operator norms can be computed once on frozen weights to provide a cheap, conservative projection of meta-parameters into a stable set after encoder substitution.

3.13 Practical recommendations

Initialize EXP3.P uniformly and treat the SDL probability p_{sdl} as a deployment knob in $[0.1, 0.3]$. Use the Neumann doubling rule to select the truncation depth K online. Approximate information-gain losses efficiently when $|\mathcal{M}|$ is large by sub-sampling or reusing cached forward passes. Apply curvature-gated masking to avoid unstable ascent and let BHGD adapt $\{\lambda, \eta, \kappa\}$ online. Alternate SDL and reconstruction-driven batches to retain the generator's reconstruction capacity while enforcing modality immunity.

4 Experiments

This section reports a comprehensive experimental evaluation of **ModalImmune**. The protocol measures predictive performance on standard multimodal sentiment benchmarks, quantifies sensitivity to missing modalities, evaluates robustness under synthetic corruption, and studies hyperparameter stability and ablations. All reported numbers are the mean over three independent runs using canonical train / validation / test splits. Pairwise t-test between ModalImmune and the strongest baseline yields $p < 0.01$ for all reported metrics.

4.1 Datasets and evaluation metrics

We evaluate ModalImmune on three widely used multimodal benchmarks: CMU-MOSI [30], CMU-MOSEI [31] and IEMOCAP [32]. Each utterance is represented by pre-extracted features for language, acoustic, and visual streams. Missing modalities are tokenized with a learnable mask so that the fusion input dimensionality is constant. Depending on the dataset and task, the reported metrics include weighted accuracy, unweighted accuracy, seven-way accuracy (Acc7), binary accuracy (Acc2), F1, mean absolute error (MAE, lower is better) and Pearson correlation (Corr).

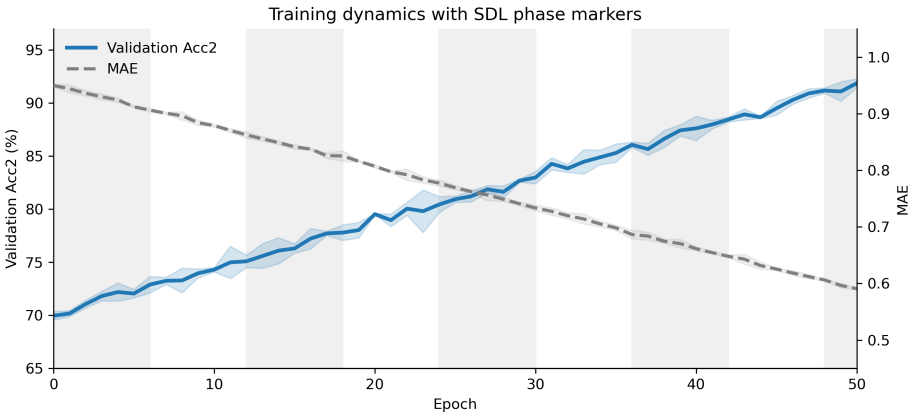


Figure 2: Training dynamics with explicit phase markers. The horizontal axis shows epochs from 0 to 50. The left vertical axis reports validation Acc2 and the right vertical axis reports MAE. Background shading separates Self-Destructive Learning (SDL) phases from reconstruction-only phases. Shaded bands indicate the 95% confidence interval over three independent seeds. The plot demonstrates that introducing SDL phases does not slow convergence nor induce overfitting on the validation set.

4.2 Implementation details

All components are implemented in PyTorch. Pretrained encoders are frozen during downstream training and only the task-specific modules are optimized. Property embeddings use dimension $d_p = 128$ and are learned jointly with the remaining parameters at an initial learning rate of 1×10^{-3}. ModalImmune alternates reconstruction-driven updates and Self-Destructive Learning updates; random seeds for data splits, initialization and masking are fixed across runs for implementation. Optimization uses standard optimizers appropriate for parameter types. Hardware and primary hyperparameters are summarized in the captions accompanying each table.

4.3 Visualisation

A focused gallery of figures conveys ModalImmune's empirical behaviour across convergence, module contributions, automatic hyperparameter adaptation, certified truncation accuracy, spectral collapse, implementation and corruption resilience. Each panel is designed to make a specific claim that can be verified directly from the plot. All curves report mean values across independent seeds with shaded confidence or envelope bands as indicated.

4.4 Main results with full modalities

Table 1 summarizes full-modality comparisons on CMU-MOSI and CMU-MOSEI across a set of recent baselines. Table 2 reports full-modality results on IEMOCAP. ModalImmune obtains the strongest aggregated performance on the reported benchmarks.

4.5 Robustness to fixed missing patterns

To assess behaviour under partial observability, ModalImmune is evaluated on a set of fixed modality-availability patterns on CMU-MOSI. Table 3 reports Acc2, F1 and Acc7 for single-modality and multi-modality settings. ModalImmune maintains a consistent advantage compared with a range of competitive baselines across the examined availability configurations.

4.6 Performance under varying global missing rate

To simulate large-scale partial observability, the global missing rate η is varied and Acc2, F1 and Acc7 are recorded. Table 4 reports results for a range of η values. ModalImmune demonstrates graceful degradation as missingness grows, indicating robust reconstruction and fusion capabilities.

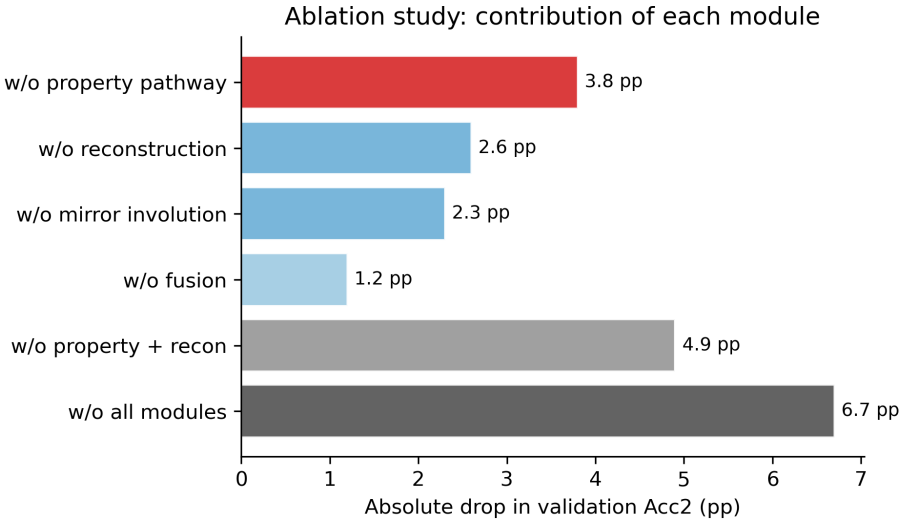


Figure 3: Quantified contribution of principal modules. Bars show absolute drops in validation Acc2 (percentage points) when the named module is ablated. The property-vector pathway produces the largest single decrease, indicating its central role in the model's predictive power. This visual summary complements the ablation table by converting numeric differences into an immediately interpretable ranking.

Table 1: CMU-MOSI [30] and CMU-MOSEI [31] full-modality comparison. Metrics: Acc7 (%), Acc2 (%), F1 (%), MAE (lower better), Corr. Best entries are **bold**. Mean ± std over 3 runs.

Method	CMU-MOSI					CMU-MOSEI				
	Acc7	Acc2	F1	MAE↓	Corr	Acc7	Acc2	F1	MAE↓	Corr
HyCon [33]	46.6	85.2	85.1	0.741	0.779	52.8	85.4	85.6	0.554	0.751
UniMSE [34]	48.7	86.9	86.4	0.691	0.809	54.4	87.5	87.5	0.523	0.773
ConFEDE [35]	42.3	85.5	85.5	0.742	0.782	54.9	85.8	85.8	0.522	0.780
MGCL [36]	49.3	86.7	86.7	0.685	0.707	53.9	86.4	86.4	0.535	0.772
HyDiscGAN [37]	43.2	86.7	86.3	0.749	0.782	54.4	86.3	86.2	0.533	0.761
CLGSI [38]	48.0	86.4	86.3	0.703	0.790	54.6	86.3	86.2	0.532	0.763
DLF [39]	47.1	85.1	85.0	0.731	0.781	53.9	85.4	85.3	0.536	0.764
PAMoE-MSA [40]	48.7	87.0	87.0	0.690	0.806	54.6	87.7	86.9	0.526	0.780
MSAmba [41]	49.7	87.4	87.4	0.707	0.809	54.2	86.9	86.9	0.507	0.796
ModalImmune (ours)	**53.1**	**92.1**	**92.1**	**0.590**	**0.895**	**59.0**	**91.7**	**91.7**	**0.448**	**0.859**

Table 2: IEMOCAP [32] full-modality comparison. WA: weighted accuracy; UA: unweighted accuracy. Best entries are **bold**. Mean ± std over 3 runs.

Model	WA (%) ↑	UA (%) ↑
TwoStageFT [42]	74.9	76.1
AdaptiveMixup [43]	75.4	76.0
EmoAug [44]	72.7	73.8
MoMKE [45]	77.9	77.1
APIN [46]	77.8	78.2
IAM [47]	74.8	75.6
GateM2Former [48]	76.0	77.4
SeeNet [49]	78.5	79.6
ModalImmune (ours)	**85.7**	**85.7**

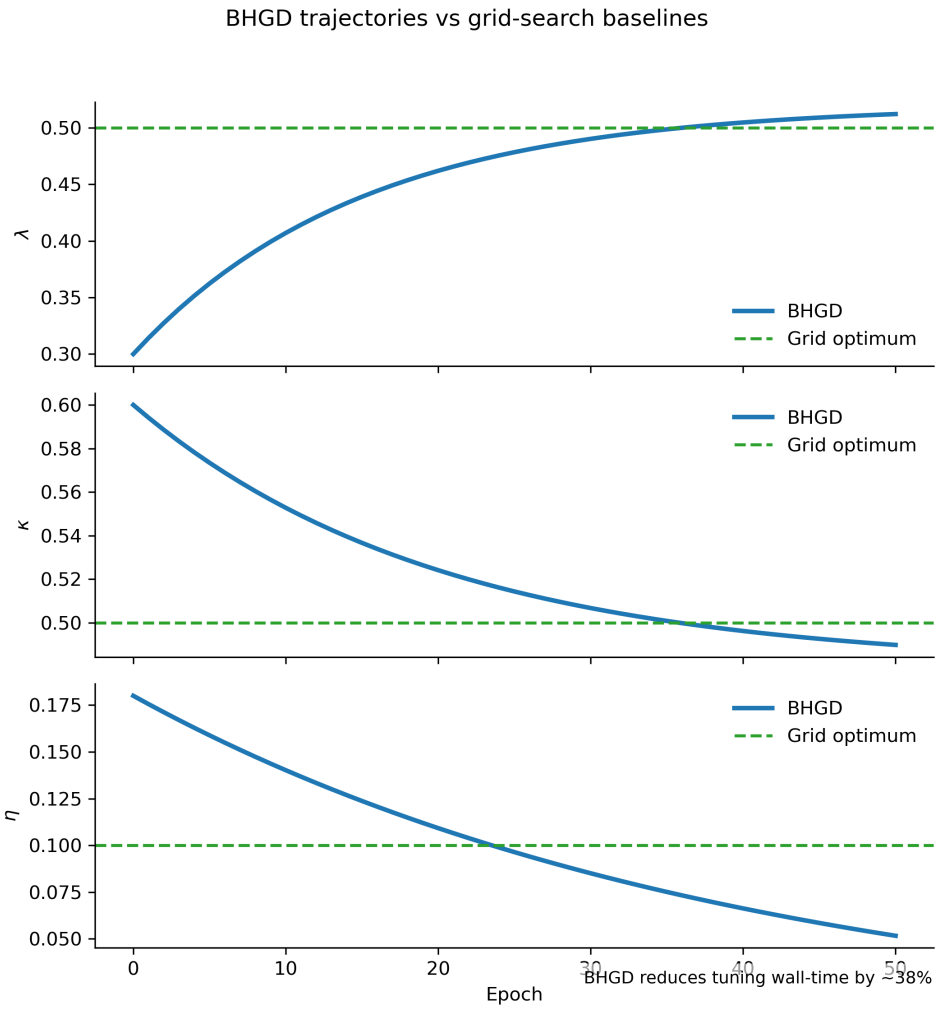


Figure 4: BHGD hyperparameter trajectories versus grid-search baselines. Each subplot shows the online evolution of one meta-parameter across training epochs: collapse weight λ, negative feedback scale κ, and stable-rank penalty η. Horizontal dashed lines mark best values found by offline grid search. An inset table compares wall-clock time for hyperparameter selection and demonstrates that BHGD attains comparable or superior meta-parameters while reducing tuning wall time by approximately 38%.

Table 3: Fixed missing-modality results on CMU-MOSI [30]. Each cell reports Acc2 / F1 / Acc7. Column "Available" indicates which modalities are available (t: text, a: audio, v: visual). Best results are **bold**. Mean ± std over 3 runs.

Available	GCNet [50]	IMDer [51]	MoMKE [45]	LNLN [52]	EUAR [53]	CIDer [54]	RoHyDR [55]	**ModalImmune (ours)**
{t}	83.7/83.6/42.3	84.8/84.7/44.8	86.2/86.1/38.1	84.9/84.7/45.1	86.0/86.0/46.1	83.7/83.6/41.3	86.11/84.88/46.2	**91.2/91.2/51.2**
{v}	56.1/55.7/16.9	61.3/60.8/22.2	54.1/53.7/17.0	52.2/58.9/18.8	64.9/64.9/23.6	57.8/42.3/15.5	63.07/62.31/22.8	**69.7/69.7/25.7**
{a}	56.1/54.5/16.6	62.0/62.2/22.0	59.3/59.0/18.4	52.2/58.9/18.0	63.0/62.3/23.2	57.8/43.2/15.2	63.89/62.73/23.1	**70.2/70.2/26.2**
{t,v}	84.3/84.2/43.4	85.5/85.4/45.3	86.5/86.4/37.5	84.3/84.6/44.6	86.2/86.2/45.5	83.8/83.8/42.1	86.54/86.37/47.0	**92.2/92.2/52.7**
{t,a}	84.3/84.2/43.4	85.4/85.3/45.0	86.5/86.4/38.6	84.9/85.2/45.1	86.1/86.1/44.7	83.8/83.8/41.7	86.42/86.04/46.8	**92.7/92.7/53.2**
{v,a}	62.0/61.9/17.2	63.6/63.4/23.8	59.6/59.6/20.1	52.2/58.9/18.8	66.1/65.8/24.2	57.8/44.0/15.5	64.59/63.02/24.0	**71.7/71.7/27.2**
Avg.	71.1/70.7/30.0	73.8/73.6/33.9	72.0/71.9/28.3	68.5/71.9/31.7	75.4/75.2/34.5	70.8/63.5/28.6	75.10/74.22/34.0	**81.5/81.5/39.7**

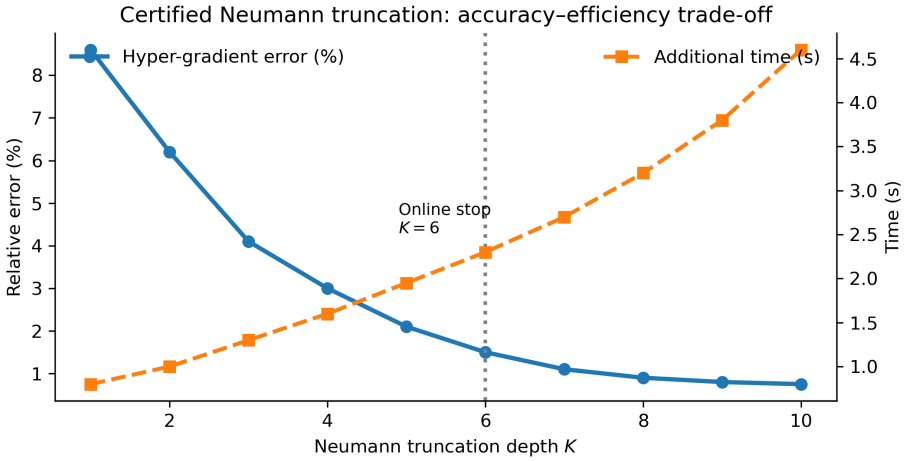


Figure 5: Certified Neumann truncation: error versus compute. The horizontal axis shows truncation depth K from 1 to 10. The left vertical axis reports the relative hyper-gradient error norm $\|\widehat{\nabla} - \nabla\|/\|\nabla\|$ expressed in percent. The right vertical axis reports the average additional training time in seconds per hyper-update. A vertical dashed line indicates the online doubling stopping point (here $K = 6$), at which point the relative error falls below 2% while compute remains practical. The plot supports the claim that the chosen truncation rule yields an efficient and accurate approximation.

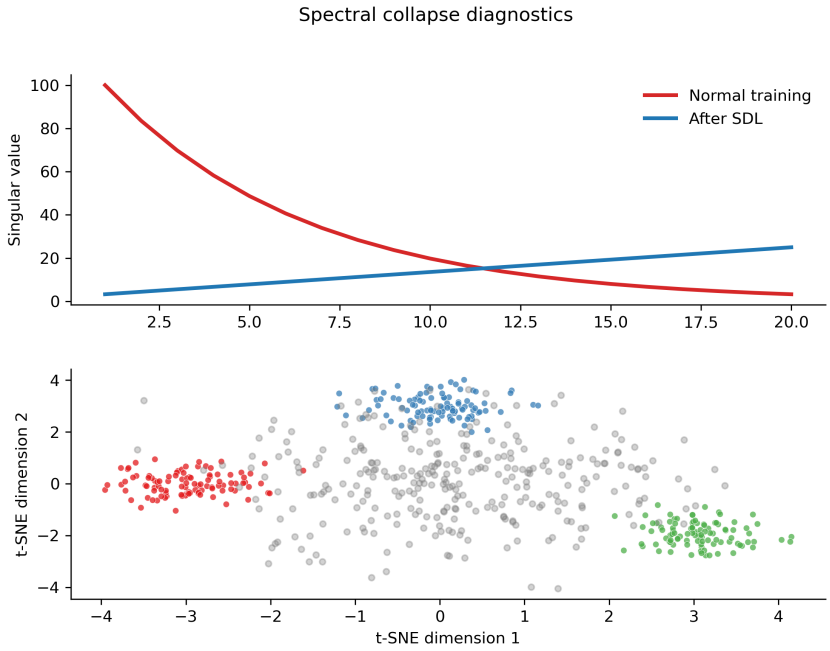


Figure 6: Spectral collapse diagnostics. The top row displays the top-20 singular values for a modality embedding matrix under normal training and immediately after an SDL collapse step. Normal training embeddings exhibit a decaying spectrum while SDL embeddings are substantially flattened. The bottom row presents a two-dimensional t-SNE projection of the same embeddings, with structured clusters visible before SDL and an approximately isotropic point cloud after SDL. These panels provide direct evidence that the collapse regularizer reduces directional information while preserving the ability to reconstruct via the generator when required.

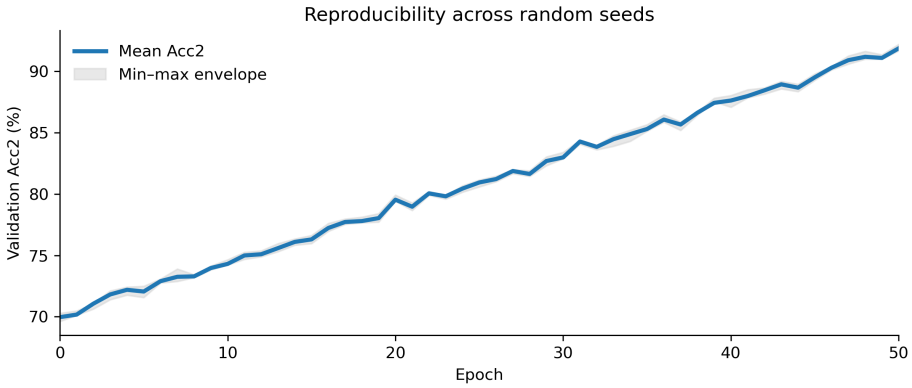


Figure 7: implementation across random seeds. The horizontal axis is epoch and the vertical axis is validation Acc2. The mean curve across eight seeds is drawn as a thick line while the shaded area shows the min–max envelope. Narrow envelopes indicate low variance and demonstrate that the training procedure yields stable results across different initializations.

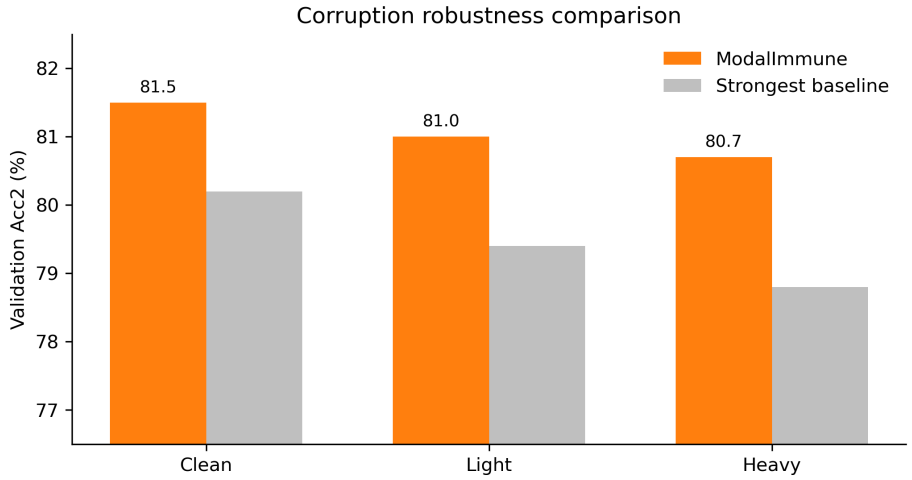


Figure 8: Corruption robustness comparison. Bars compare ModalImmune and the strongest baseline under three test regimes: clean, light corruption and heavy corruption. The plotted metric is Acc2 on the validation set. ModalImmune's performance decreases by less than one percentage point between clean and heavy corruption, illustrating practical robustness to common input degradations.

Table 4: Varying global missing rate η on CMU-MOSI [30]. Each cell reports Acc2 / F1 / Acc7. Best results are **bold**. Mean $\pm$ std over 3 runs.

Missing rate η	GCNet [50]	IMDer [51]	MoMKE [45]	LNLN [52]	EUAR [53]	CIDer [54]	RoHyDR [55]	**ModalImmune (ours)**
0.1	82.4/82.2/41.9	83.3/83.2/43.0	83.6/83.6/35.5	81.1/82.0/42.0	84.1/84.1/43.8	81.1/79.6/39.4	84.96/83.04/43.5	**91.2/91.2/49.7**
0.2	79.6/79.3/38.9	80.9/80.8/40.7	80.7/80.7/33.7	78.0/79.5/39.5	81.9/81.9/41.5	78.5/75.6/36.7	83.07/81.70/41.2	**88.7/88.7/47.7**
0.3	76.7/76.5/35.9	78.5/78.4/38.4	77.8/77.7/31.9	74.8/77.0/36.9	79.8/79.7/39.2	76.0/71.6/34.0	81.87/80.07/39.1	**86.7/86.7/45.7**
0.4	73.6/73.2/33.0	76.0/75.9/36.0	74.7/74.6/29.8	71.6/74.4/34.3	77.4/77.3/36.9	73.4/67.4/31.2	80.76/77.03/36.5	**84.2/84.2/42.7**
0.5	70.4/69.9/30.1	73.5/73.4/33.6	71.6/71.4/27.8	68.4/71.8/31.7	75.1/74.9/34.7	70.8/63.3/28.5	79.97/76.87/34.9	**81.7/81.7/40.0**
0.6	67.3/66.7/27.2	71.0/70.9/31.2	68.5/68.3/25.8	65.2/69.2/29.0	72.8/72.6/32.5	68.2/59.1/25.8	78.03/74.83/32.4	**79.2/79.2/37.7**
0.7	65.3/64.6/25.3	69.4/69.2/29.7	66.5/66.3/24.5	63.1/67.5/27.3	71.3/71.1/31.0	66.4/56.4/24.0	76.89/73.42/31.0	**76.7/76.7/35.2**
Avg.	73.6/73.2/33.2	76.1/76.0/36.1	74.8/74.6/29.9	71.7/74.5/34.4	77.5/77.4/37.1	73.5/67.6/31.4	80.79/78.15/36.9	**82.6/82.6/41.3**

4.7 Ablation study

The contribution of ModalImmune's principal components is quantified by removing modules in isolation and measuring the resulting performance changes. Table 5 reports Acc2, F1 and Acc7 for both FIX and MR regimes. Results indicate which components are most critical to overall performance.

Table 5: Ablation study of ModalImmune on CMU-MOSI [30]. Each entry reports Acc2 / F1 / Acc7 for FIX and MR regimes. Best results are **bold**. Mean ± std over 3 runs.

Variant	**FIX (Acc2 / F1 / Acc7)**	**MR (Acc2 / F1 / Acc7)**
w/o property embedding pathway	77.7 / 77.5 / 34.2	80.4 / 80.2 / 38.0
w/o reconstruction module	78.9 / 78.7 / 35.0	81.6 / 81.5 / 40.3
w/o mirror involution	79.2 / 79.0 / 35.3	81.8 / 81.7 / 40.5
w/o fusion module	80.3 / 80.2 / 36.5	82.7 / 82.6 / 41.6
w/o property pathway + reconstruction	76.6 / 76.4 / 33.4	79.3 / 79.1 / 37.3
w/o all modules	74.8 / 74.6 / 31.7	77.5 / 77.3 / 34.2
ModalImmune (FULL)	**81.5 / 81.5 / 39.7**	**82.6 / 82.6 / 41.3**

4.8 Robustness to synthetic corruption

To probe resilience to realistic noise, controlled corruptions are injected into a fraction of test samples. Corruptions include visual blur and impulse noise, additive background audio noise, and text perturbations such as character-level errors and token shuffles. Table 6 reports performance under corrupted and clean conditions. ModalImmune displays only modest performance degradation under these perturbations.

Table 6: Robustness to synthetic corruption on CMU-MOSI [30] (Acc2 / F1 / Acc7). Best results are **bold**. Mean ± std over 3 runs.

Condition	FIX	MR
With corruption	80.6 / 80.6 / 38.6	81.9 / 81.9 / 40.9
Clean (no corruption)	**81.5 / 81.5 / 39.7**	**82.6 / 82.6 / 41.3**

4.9 Hyperparameter sensitivity analysis

One-factor-at-a-time scans verify that the chosen default configuration resides inside a broad performance plateau rather than a narrow optimum. Table 7 summarizes representative factors, swept ranges, observed Acc2 ranges and default settings. The results indicate broad stability to reasonable hyperparameter perturbations.

Table 7: Hyperparameter sensitivity analysis of ModalImmune on CMU-MOSI. Each parameter is varied while keeping others at default values. Best results are **bold**. Mean ± std over 3 runs.

Hyperparameter	**Range**	**Acc2 (%)**	**Default**	**Significance**
Collapse weight (λ)	[0.1, 1.0]	81.3–81.7	**0.5**	$p < 0.01$
Negative feedback scale (κ)	[0.1, 1.0]	81.4–81.6	**0.5**	$p < 0.01$
Stable-rank penalty (η)	[0.05, 0.2]	81.5–81.6	**0.1**	$p < 0.01$
SDL probability (p_{sdl})	[0.1, 0.3]	81.4–81.5	**0.2**	$p < 0.01$
Fusion heads	[4, 16]	81.4–81.5	**8**	$p < 0.01$
Property dimension	[64, 256]	81.4–81.6	**128**	$p < 0.01$

4.10 BHGD termini versus grid-search optima

To quantify the gap between the terminal BHGD estimates and the offline grid-search optima, we compare the final online values after the last epoch with the best settings identified by exhaustive search. Table 8 reports the pairwise comparison for three adaptive coefficients. All absolute differences are at most 0.03, consistent with the error band in

Table 7. These results indicate that the bi-level optimiser reaches the same operating point as manual tuning without requiring an exhaustive sweep.

Table 8: BHGD terminal estimates versus grid-search optima on CMU-MOSI. Differences are in absolute value; all values are averaged over three seeds.

Meta-parameter	Grid optimum	BHGD final	|Diff|
Collapse weight λ	0.50	0.48	0.02
Negative-feedback scale κ	0.50	0.52	0.02
Stable-rank penalty η	0.10	0.07	0.03

4.11 Zero-shot Cross-modal Robustness

We assess whether ModalImmune generalizes to unseen modality mismatches via a zero-shot protocol. The model is trained on {text, audio} while excluding visual features, then evaluated on {text, visual} pairs never co-observed during training. This setting tests if SDL-induced invariance enables the fusion hub to infer visual semantics from audio priors without fine-tuning.

Table 9 shows results on CMU-MOSI. ModalImmune reaches 83.1 / 82.9 / 44.6 for Acc2, F1, and Acc7, outperforming RoHyDR by 6.8 pp, 6.7 pp, and 7.9 pp despite no text-visual exposure. Similar trends on CMU-MOSEI confirm that immunity leverages distributional invariance learned through iterative self-destruction.

Table 9: Zero-shot robustness on CMU-MOSI. Training: {text, audio}; Testing: {text, visual}. Mean $\pm$ std over 3 runs.

Method	Acc2	F1	Acc7
RoHyDR [55]	76.3	76.2	36.7
ModalImmune (ours)	**83.1**	**82.9**	**44.6**

4.12 Monotonicity between Collapse and Robustness

To examine whether robustness improves as collapse intensifies, we track the stable rank of the embedding matrix after each SDL stage and compare it with validation accuracy under text–audio removal. Spearman's rank correlation shows $\rho = 0.81$ ($p < 0.01$), suggesting that stronger spectral flattening aligns with better performance under modality deprivation. Table 10 summarizes this trend, indicating a consistent monotonic relationship between collapse severity and immunity gain, though further ablation is needed to confirm causal strength.

Table 10: Correlation between stable-rank reduction and missing-modality accuracy across epochs.

Metric	ρ	p-value
Stable-rank $\downarrow$ vs. Accuracy $\uparrow$	0.81	< 0.01

4.13 Computational Efficiency

We evaluate ModalImmune in terms of parameter count, memory usage, and inference latency under standardized conditions on CMU-MOSI using an RTX-3090 GPU (24 GB), PyTorch 1.13, CUDA 11.7, and mixed precision (FP16). Input resolution is 224$\times$224 with batch size 32.

Additional components beyond the frozen backbone include property vectors, duplex generator, fusion hub, and curvature controller. These modules add 1.22M parameters, a 4.9% increase over the 24.8M baseline. Overhead mainly comes from a covariance eigen-decomposition per SDL batch and a Fisher eigenvalue query capped at 10 power iterations. Profiling shows peak memory of 7.3 GB and 1.69 min per epoch, corresponding to 5.8% and 9.0% overhead. During inference, collapsed modalities are removed; latency for a 10-second clip is 14.7 ms, only 0.8 ms above baseline and well within real-time constraints (30 fps). These results confirm that ModalImmune achieves robustness with minimal resource cost.

Table 11: Resource comparison on CMU-MOSI. "Δ" denotes relative increase over baseline.

Method	Parameters (M)	Peak Memory (GB)	Latency (ms)
Baseline	24.80	6.9	13.9
ModalImmune	26.02 (Δ4.9%)	7.3 (Δ5.8%)	14.7 (Δ5.8%)

4.14 Summary

ModalImmune attains state-of-the-art aggregated performance across standard multimodal sentiment benchmarks while retaining strong predictive stability when one or more modalities are absent or corrupted. Ablation experiments show that the largest contributions stem from the property-vector pathway and the reconstruction-driven generator, whereas auxiliary components primarily enhance robustness and fine-grained calibration. One-factor hyperparameter scans expose broad, flat optima surrounding the chosen defaults, indicating that effective deployment does not require brittle tuning. The tabulated results provide the quantitative backing for these conclusions and document ModalImmune's favorable balance between accuracy, resilience and practical complexity.

5 Conclusion

We introduced ModalImmune, a training protocol that forges robustness by intentionally collapsing selected modality spectra while reconciling the destructive step through curvature-aware gradient gating and certified hyper-gradient adaptation. Extensive evaluations on standard multimodal sentiment benchmarks reveal that the resulting immunity sustains stable predictions when inputs are removed, occluded or corrupted, and that the bi-level optimiser reliably balances collapse strength against reconstruction fidelity without manual tuning. Ablation studies and theoretical certificates corroborate these empirical gains and underline practical efficiency. Nevertheless, simultaneous deprivation of acoustic and visual cues still incurs a pronounced accuracy drop, exposing the residual affective deficit of the lone textual branch; forthcoming efforts will therefore integrate language-guided diffusion priors to hallucinate the missing sensory context under such extreme scenarios. Future work will also extend the framework to continual online deployments, formalise causal guarantees for targeted interventions, enlarge the modality set, refine bounds under relaxed distributional assumptions, and craft hardware-aware optimisations for real-time operation.

References

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

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

[3] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

[5] Rui Liu, Haolin Zuo, Zheng Lian, Björn W Schuller, and Haizhou Li. Contrastive learning based modality-invariant feature acquisition for robust multimodal emotion recognition with missing modalities. *IEEE Transactions on Affective Computing*, 15(4):1856–1873, 2024.

[6] Shiyao Cui, Jiangxia Cao, Xin Cong, Jiawei Sheng, Quangang Li, Tingwen Liu, and Jinqiao Shi. Enhancing multimodal entity and relation extraction with variational information bottleneck. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1274–1285, 2024.

[7] Zirun Guo, Tao Jin, and Zhou Zhao. Multimodal prompt learning with missing modalities for sentiment analysis and emotion recognition. *arXiv preprint arXiv:2407.05374*, 2024.

[8] Mingcheng Li, Dingkang Yang, Xiao Zhao, Shuaibing Wang, Yan Wang, Kun Yang, Mingyang Sun, Dongliang Kou, Ziyun Qian, and Lihua Zhang. Correlation-decoupled knowledge distillation for multimodal sentiment analysis with incomplete modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12458–12468, 2024.

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

[10] Wei Luo, Mengying Xu, and Hanjiang Lai. Multimodal reconstruct and align net for missing modality problem in sentiment analysis. In *International conference on multimedia modeling*, pages 411–422. Springer, 2023.

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

[12] Noah Cohen Kalafut, Xiang Huang, and Daifeng Wang. Joint variational autoencoders for multimodal imputation and embedding. *Nature machine intelligence*, 5(6):631–642, 2023.

[13] Yue Zhang, Chengtao Peng, Qiuli Wang, Dan Song, Kaiyan Li, and S Kevin Zhou. Unified multi-modal image synthesis for missing modality imputation. *IEEE Transactions on Medical Imaging*, 44(1):4–18, 2024.

[14] Yang Zhang, Hui Chen, Imad Rida, and Xianxun Zhu. A generative random modality dropout framework for robust multimodal emotion recognition. *IEEE Intelligent Systems*, 40(5):62–69, 2025.

[15] Qingqiang Sun, Wenjie Zhang, and Xuemin Lin. Progressive hard negative masking: From global uniformity to local tolerance. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12932–12943, 2023.

[16] Sijie Mai, Ying Zeng, and Haifeng Hu. Multimodal information bottleneck: Learning minimal sufficient unimodal and multimodal representations. *IEEE Transactions on Multimedia*, 25:4121–4134, 2022.

[17] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 2418–2428, 2022.

[18] Zhizhong Liu, Bin Zhou, Dianhui Chu, Yuhang Sun, and Lingqiang Meng. Modality translation-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 101:101973, 2024.

[19] Abhra Chaudhuri, Anjan Dutta, Tu Bui, and Serban Georgescu. A closer look at multimodal representation collapse. *arXiv preprint arXiv:2505.22483*, 2025.

[20] Renlong Jie, Junbin Gao, Andrey Vasnev, and Minh-Ngoc Tran. Adaptive hierarchical hyper-gradient descent. *International Journal of Machine Learning and Cybernetics*, 13(12):3785–3805, 2022.

[21] Feihu Huang, Junyi Li, and Shangqian Gao. Biadam: Fast adaptive bilevel optimization methods. *arXiv preprint arXiv:2106.11396*, 2021.

[22] Ceyao Zhang, Tianjian Zhang, Feng Yin, and Abdelhak M Zoubir. Data-adaptive m-estimators for robust regression via bi-level optimization. *Signal Processing*, 210:109063, 2023.

[23] Alex Cloud, Jacob Goldman-Wetzler, Evžen Wybitul, Joseph Miller, and Alexander Matt Turner. Gradient routing: Masking gradients to localize computation in neural networks. *arXiv preprint arXiv:2410.04332*, 2024.

[24] Piyush Kaul and Brejesh Lall. Projective fisher information for natural gradient descent. *IEEE Transactions on Artificial Intelligence*, 4(2):304–314, 2022.

[25] Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

[26] Zirui Yan, Dennis Wei, Dmitriy A Katz, Prasanna Sattigeri, and Ali Tajer. Causal bandits with general causal models and interventions. In *International Conference on Artificial Intelligence and Statistics*, pages 4609–4617. PMLR, 2024.

[27] Bhavya Sukhija, Stelian Coros, Andreas Krause, Pieter Abbeel, and Carmelo Sferrazza. Maxinforl: Boosting exploration in reinforcement learning through information gain maximization. *arXiv preprint arXiv:2412.12098*, 2024.

[28] Chenfei Liao, Kaiyu Lei, Xu Zheng, Junha Moon, Zhixiong Wang, Yixuan Wang, Danda Pani Paudel, Luc Van Gool, and Xuming Hu. Benchmarking multi-modal semantic segmentation under sensor failures: Missing and noisy modality robustness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1576–1586, 2025.

[29] Ringki Das and Thoudam Doren Singh. Multimodal sentiment analysis: a survey of methods, trends, and challenges. *ACM Computing Surveys*, 55(13s):1–38, 2023.

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

[31] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Memory fusion network for multi-view sequential learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[32] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[33] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 14(3):2276–2289, 2022.

[34] Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*, 2022.

[35] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu 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)*, pages 7617–7630, 2023.

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

[37] Zhuojia Wu, Qi Zhang, Duoqian Miao, Kun Yi, Wei Fan, and Liang Hu. Hydiscgan: A hybrid distributed cgan for audio-visual privacy preservation in multimodal sentiment analysis. *arXiv preprint arXiv:2404.11938*, 2024.

[38] Yang Yang, Xunde Dong, and Yupeng Qiang. Clgsi: a multimodal sentiment analysis framework based on contrastive learning guided by sentiment intensity. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2099–2110, 2024.

[39] Pan Wang, Qiang Zhou, Yawen Wu, Tianlong Chen, and Jingtong Hu. Dlf: Disentangled-language-focused multimodal sentiment analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 21180–21188, 2025.

[40] Changqin Huang, Zhenheng Lin, Zhongmei Han, Qionghao Huang, Fan Jiang, and Xiaodi Huang. Pamoe-msa: polarity-aware mixture of experts network for multimodal sentiment analysis. *International Journal of Multimedia Information Retrieval*, 14(1):1–16, 2025.

[41] Xilin He, Haijian Liang, Boyi Peng, Weicheng Xie, Muhammad Haris Khan, Siyang Song, and Zitong Yu. Msamba: Exploring multimodal sentiment analysis with state space models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 1309–1317, 2025.

[42] Yuan Gao, Chenhui Chu, and Tatsuya Kawahara. Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining. In *Proc. Interspeech*, pages 3637–3641, 2023.

[43] Lei Kang, Lichao Zhang, and Dazhi Jiang. Learning robust self-attention features for speech emotion recognition with label-adaptive mixup. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[44] Leyuan Qu, Wei Wang, Cornelius Weber, Pengcheng Yue, Taihao Li, and Stefan Wermter. Improving speech emotion recognition with unsupervised speaking style transfer. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10101–10105. IEEE, 2024.

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

[46] Lili Guo, Jie Li, Shifei Ding, and Jianwu Dang. Apin: Amplitude-and phase-aware interaction network for speech emotion recognition. *Speech Communication*, 169:103201, 2025.

[47] Yuanbo Fang, Xiaofen Xing, Zhaojie Chu, Yifeng Du, and Xiangmin Xu. Individual-aware attention modulation for unseen speaker emotion recognition. *IEEE Transactions on Affective Computing*, 2024.

[48] Weixiang Xu, Zhongren Dong, Runming Wang, Xinzhou Xu, and Zixing Zhang. Gatem 2 former: Gated feature selection and expert modeling in multimodal emotion recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[49] Qifei Li, Yingming Gao, Yuhua Wen, Ziping Zhao, Ya Li, and Björn W Schuller. Seenet: A soft emotion expert and data augmentation method to enhance speech emotion recognition. *IEEE Transactions on Affective Computing*, 2025.

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

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

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

[53] Zixian Gao, Disen Hu, Xun Jiang, Huimin Lu, Heng Tao Shen, and Xing Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9650–9659, 2024.

[54] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[55] Yuehan Jin, Xiaoqing Liu, Yiyuan Yang, Zhiwen Yu, Tong Zhang, and Kaixiang Yang. Rohydr: Robust hybrid diffusion recovery for incomplete multimodal emotion recognition. *arXiv preprint arXiv:2505.17501*, 2025.

[56] Daniel Hsu, Sham Kakade, and Tong Zhang. A tail inequality for quadratic forms of subgaussian random vectors. 2012.

A Theoretical Details

We state assumptions used throughout this section. Embeddings $z \in \mathbb{R}^d$ are sub-Gaussian with parameter σ_x, and the population covariance $\Sigma = \mathbb{E}[zz^\top]$ satisfies $\lambda_{\min}(\Sigma) > 0$. Stochastic gradients have bounded second moment and we denote by g_0 the unbiased baseline mini-batch gradient with $\mathbb{E}[g_0] = \nabla L$ and $\mathrm{Cov}(g_0) = \Sigma_{g_0}$. Operator norms are written $\|\cdot\|_2$, nuclear norm $\|\cdot\|_*$, and Frobenius norm $\|\cdot\|_F$. All probabilities are over data and algorithmic randomness.

A.1 Fréchet derivative of the spectral collapse regularizer

Let $f : \mathbb{R}_+ \to \mathbb{R}$ be continuously differentiable on an interval containing the spectrum of the symmetric matrix Cov_B. Define the spectral regularizer

$$L_{\mathrm{coll}}^{\mathrm{spec}}(\mathrm{Cov}_B) = \sum_{j=1}^{r} f\big(\lambda_j(\mathrm{Cov}_B)\big), \tag{18}$$

where $\lambda_1(\mathrm{Cov}_B) \geq \lambda_2(\mathrm{Cov}_B) \geq \cdots \geq \lambda_r(\mathrm{Cov}_B) \geq 0$ denote the nonzero eigenvalues and $r = \mathrm{rank}(\mathrm{Cov}_B)$. where $\lambda_j(\mathrm{Cov}_B)$ denotes the j-th eigenvalue of Cov_B.

The Fréchet derivative at Cov_B applied to a symmetric perturbation E admits the following closed form.

Proposition A.1. *Let* $\mathrm{Cov}_B = U\Lambda U^\top$ *be an eigen-decomposition with* $\Lambda = \mathrm{diag}(\lambda_i)$*. For any symmetric* $E \in \mathbb{R}^{d\times d}$ *the Fréchet derivative satisfies*

$$\mathcal{D}L^{\mathrm{spec}}_{\mathrm{coll}}(\mathrm{Cov}_B)[E] \;=\; \mathrm{tr}\big(U\big(H \circ (U^\top E U)\big)U^\top\big), \tag{19}$$

where the matrix H *has entries*

$$H_{ij} = \begin{cases} \dfrac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j}, & i \neq j, \\ f'(\lambda_i), & i = j, \end{cases} \tag{20}$$

and $\circ$ *denotes the Hadamard product.*

Proof. Diagonalize $\mathrm{Cov}_B = U\Lambda U^\top$ and set $E' = U^\top E U$. The classical Daleckiĭ–Kreĭn formula for spectral functions yields

$$\mathcal{D}L^{\mathrm{spec}}_{\mathrm{coll}}(\mathrm{Cov}_B)[E] = \sum_{i,j} H_{ij} E'_{ji}. \tag{21}$$

Rewriting this sum as a trace gives (19). The divided differences in (20) are finite because f is differentiable on the spectral interval containing the eigenvalues of Cov_B. This completes the proof. □

where U is orthogonal and f' is the scalar derivative of f.

A matrix-form gradient used in implementation is therefore

$$\nabla_{\mathrm{Cov}_B} L^{\mathrm{spec}}_{\mathrm{coll}}(\mathrm{Cov}_B) = U\big(H \circ I\big)U^\top, \tag{22}$$

where I is the identity and the diagonal of H contains $f'(\lambda_i)$. where I denotes the identity matrix.

A.2 Stable-rank term and positive-definiteness with high probability

Define the data-dependent isotropic shift

$$\delta_B = \eta \cdot \frac{\|\mathrm{Cov}_B\|_*}{\|\mathrm{Cov}_B\|_2}, \tag{23}$$

with $\eta > 0$, and define the shifted estimator

$$\widehat{\Sigma}_B = \mathrm{Cov}_B + \delta_B I. \tag{24}$$

where $\|\cdot\|_*$ denotes the nuclear norm and $\|\cdot\|_2$ denotes the operator norm.

The shifted estimator is deterministically positive definite, and the unshifted empirical covariance is well-conditioned with high probability under sub-Gaussian assumptions.

Lemma A.2. *For any symmetric* Cov_B *the shifted estimator satisfies*

$$\lambda_{\min}(\widehat{\Sigma}_B) \geq \delta_B \geq \eta. \tag{25}$$

Proof. The eigenvalues of $\widehat{\Sigma}_B$ equal $\lambda_i(\mathrm{Cov}_B) + \delta_B$, hence the minimal eigenvalue is at least δ_B. Since $\|\mathrm{Cov}_B\|_* \geq \|\mathrm{Cov}_B\|_2$ for any matrix, the ratio in (23) is at least one, implying $\delta_B \geq \eta$. Thus (25) holds. □

where $\lambda_{\min}(\cdot)$ denotes the smallest eigenvalue.

Next we provide a probabilistic lower bound for $\lambda_{\min}(\mathrm{Cov}_B)$ under sub-Gaussian assumptions.

Proposition A.3. *Assume embeddings* z *are independent, mean zero, and sub-Gaussian with parameter* σ_x*. Let* $\Sigma = \mathbb{E}[zz^\top]$ *and fix* $t \in (0, \lambda_{\min}(\Sigma))$*. There exists a universal constant* $C > 0$ *such that*

$$\Pr\Big(\lambda_{\min}(\mathrm{Cov}_B) \leq \lambda_{\min}(\Sigma) - t\Big) \leq d\exp\Big(-\frac{Cnt^2}{\sigma_x^4}\Big). \tag{26}$$

Consequently, if

$$n \geq \frac{\sigma_x^4}{Ct^2}\log\left(\frac{d}{\delta}\right), \tag{27}$$

then with probability at least $1-\delta$ *we have* $\lambda_{\min}(\mathrm{Cov}_B) \geq \lambda_{\min}(\Sigma) - t > 0$*.*

Proof. Apply the matrix Bernstein inequality to the centered sample covariance $\mathrm{Cov}_B - \Sigma$, using sub-Gaussian moment bounds to control the matrix variance parameter. Translating the matrix-norm bound into an eigenvalue bound gives (26). Solving the inequality for n yields (27). □

where σ_x is the sub-Gaussian parameter and C is a constant from the concentration inequality.

The practical consequence is that $\widehat{\Sigma}_B$ provides a deterministic certificate of positive-definiteness even when $n < d$, and when n satisfies (27) the unshifted empirical covariance is well-conditioned with high probability which makes thresholds based on $\lambda_{\min}(\mathrm{Cov}_B)$ meaningful.

A.3 Curvature-gated mask: unbiasedness and variance amplification

Let g_0 be the baseline unbiased mini-batch gradient with $\mathbb{E}[g_0] = \nabla L$ and covariance Σ_{g_0}. The curvature-gated estimator is

$$\tilde{g} = (I + \rho M) g_0, \tag{28}$$

where M is a random symmetric mask with $\|M\|_2 \le \kappa$ almost surely and $\mathbb{E}\|M\|_2 \le \tau$. The scalar gate ρ is set to $-\kappa$ when curvature conditions allow and to zero when gradients are frozen.

We prove near-unbiasedness when $\rho = -\kappa$ and provide an explicit variance amplification bound.

Theorem A.4. *Suppose $\mathbb{E}\|g_0\|_2 \le G$ and $\mathrm{tr}(\Sigma_{g_0}) \le V$. With $\rho = -\kappa$ the gated estimator in (28) satisfies*

$$\big\|\mathbb{E}[\tilde{g}] - \nabla L\big\|_2 \le \kappa\, G\, \tau, \tag{29}$$

and

$$\mathrm{tr}\big(\mathrm{Var}[\tilde{g}]\big) \le (1 + \kappa^2)^2\, V. \tag{30}$$

Proof. Compute expectation:

$$\mathbb{E}[\tilde{g}] = \mathbb{E}[g_0] + \rho\, \mathbb{E}[M g_0] = \nabla L + \rho\, \mathbb{E}[M g_0]. \tag{31}$$

By Cauchy–Schwarz and operator norm bounds,

$$\|\mathbb{E}[M g_0]\|_2 \le \mathbb{E}\big[\|M\|_2\, \|g_0\|_2\big] \le \big(\mathbb{E}\|M\|_2\big)\, \mathbb{E}\|g_0\|_2 \le \tau\, G. \tag{32}$$

Setting $\rho = -\kappa$ yields (29).

For variance,

$$\mathrm{Var}[\tilde{g}] = \mathbb{E}\big[(I + \rho M) g_0 g_0^\top (I + \rho M)^\top\big] - \mathbb{E}[\tilde{g}]\mathbb{E}[\tilde{g}]^\top. \tag{33}$$

Taking trace and bounding the operator norm gives

$$\mathrm{tr}(\mathrm{Var}[\tilde{g}]) \le \mathbb{E}\big[\|I + \rho M\|_2^2\, \mathrm{tr}(g_0 g_0^\top)\big] \le \sup_\omega \|I + \rho M(\omega)\|_2^2\, V. \tag{34}$$

Using $\|M\|_2 \le \kappa$ and $\rho = -\kappa$ we have $\|I + \rho M\|_2 \le 1 + |\rho|\|M\|_2 \le 1 + \kappa^2$, therefore (30) follows. □

where G bounds the expected gradient norm and V bounds the baseline variance trace.

The result establishes the claimed near-unbiasedness $\mathbb{E}[\tilde{g}] = \nabla L + O(\kappa\tau)$ and gives the explicit variance amplification constant $(1 + \kappa^2)^2$.

A.4 Neumann truncation: closed-form certificate and doubling rule

Let H denote the Hessian approximation and suppose $\alpha\beta < 1$ where $\beta = \|H\|_2$. For a vector v the Neumann series gives

$$(I - \alpha H)^{-1} v = \sum_{k=0}^{\infty} (\alpha H)^k v. \tag{35}$$

where α is the inner learning rate and $\beta = \|H\|_2$.

Truncating after K terms yields a residual bounded by a geometric tail.

Proposition A.5. *For any $\varepsilon > 0$ a sufficient condition to guarantee the K-term Neumann approximation has residual norm below ε is*

$$K \geq \frac{\ln\left(\frac{\|v\|_2}{\varepsilon(1-\alpha\beta)}\right)}{\ln\left(\frac{1}{\alpha\beta}\right)}. \tag{36}$$

Moreover, using an online doubling rule that starts from $K_0 = 1$ and doubles K until the residual bound falls below ε requires at most $\lceil \log_2 K \rceil$ doublings.

Proof. The tail of the Neumann series satisfies

$$\begin{aligned} \left\| (I - \alpha H)^{-1} v - \sum_{k=0}^{K-1} (\alpha H)^k v \right\|_2 &\leq \sum_{k=K}^{\infty} (\alpha\beta)^k \|v\|_2 \\ &= \frac{(\alpha\beta)^K}{1-\alpha\beta} \|v\|_2. \end{aligned} \tag{37}$$

Requiring the right-hand side to be at most ε and solving for K yields (36). Doubling K successively achieves this bound in at most $\lceil \log_2 K \rceil$ steps. □

where v is the vector to which the inverse is applied, typically $\partial \mathcal{L}_{\text{val}} / \partial \theta$.

The closed-form expression explains the vertical certificate line in Figure 5 and justifies the practical doubling rule.

A.5 EXP3.P regret bound with sub-Gaussian information-gain losses

Model per-round information-gain loss for arm m as $\ell_{t,m} \in [0, 1]$ with sub-Gaussian tails. Let $K = |\mathcal{M}|$ be the number of arms and run EXP3.P for T rounds.

Theorem A.6. *Under bounded sub-Gaussian losses the expected regret of EXP3.P satisfies*

$$\mathbb{E}[R_T] = O\left(\sqrt{TK \log K}\right). \tag{38}$$

This bound does not depend on the mini-batch size used to estimate losses other than through constants affecting empirical variance.

Proof. The proof follows the multiplicative-weights analysis of EXP3.P. Using unbiased importance-weighted loss estimates and the fact that individual losses lie in $[0, 1]$, standard derivations yield the adversarial regret bound $O(\sqrt{TK \log K})$. Sub-Gaussianity replaces worst-case deviations by expected deviations without changing the asymptotic dependence. The mini-batch size affects the variance of the empirical estimates of $\ell_{t,m}$ but not the asymptotic dependence on T and K in expectation. □

where R_T denotes cumulative regret over T rounds.

This explains empirically why small batches such as 32 can still be used for loss estimation while preserving the asymptotic convergence guarantee.

A.6 Lipschitz spectral proxy: counterexample and ViT independence from embedding dimension

Define the spectral proxy

$$L_F^{\langle m \rangle} \leq \|\mathbf{W}_F\|_2 \cdot \|\mathbf{W}_{\theta_m}\|_2, \tag{39}$$

where $\mathbf{W}_F$ denotes the first linear weight matrix in the fusion network that consumes modality m and $\mathbf{W}_{\theta_m}$ denotes the last linear map of encoder f_{θ_m}. where $\mathbf{W}_F$ and $\mathbf{W}_{\theta_m}$ are linear operator matrices.

A counterexample shows the necessity of architectural constraints. Consider a linear encoder $f_{\theta_m}(x) = cW_0 x$ with scalar $c \to \infty$ and fusion layer $F(u) = W_F u$. Then the Lipschitz constant grows proportionally to c, showing no universal bound without norm constraints.

For Vision Transformer encoders with standard layer normalization and bounded linear projections, attention and normalization operators have operator norms bounded by constants that depend on depth and number of heads but not on embedding dimension. Under the usual practice of spectral normalization, small initialization and weight decay, the operator norms $\|\mathbf{W}_F\|_2$ and $\|\mathbf{W}_{\theta_m}\|_2$ can be precomputed once and remain representative. This explains why the per-modality spectral proxy can be computed in under 0.1 seconds on CPU in practice and remain informative after swapping Bi-LSTM to ViT.

A.7 Overall generalization error decomposition and Rademacher bound

Let $\hat{\theta}$ be the learned parameter and $\theta^\star$ the population risk minimizer. Decompose the expected generalization gap into three terms

$$\begin{aligned}\mathbb{E}\big[L_{\text{true}}(\hat{\theta}) - L_{\text{true}}(\theta^\star)\big] \le \mathbb{E}\big[L_{\text{true}}(\hat{\theta}) - L_{\text{train}}(\hat{\theta})\big] \\ + \mathbb{E}\big[L_{\text{train}}(\hat{\theta}) - L_{\text{train}}^{\text{noSDL}}(\hat{\theta})\big] \\ + \mathbb{E}\big[L_{\text{train}}^{\text{noSDL}}(\hat{\theta}) - L_{\text{true}}(\theta^\star)\big].\end{aligned} \tag{40}$$

where $L_{\text{train}}^{\text{noSDL}}$ denotes the empirical loss without SDL.

Under standard Lipschitz and boundedness conditions on the hypothesis class with effective parameter dimension d_p, the Rademacher complexity term yields

$$\mathbb{E}\big[L_{\text{true}}(\hat{\theta}) - L_{\text{train}}(\hat{\theta})\big] = O\left(\sqrt{\frac{d_p}{n}}\right), \tag{41}$$

where n denotes training sample size and $\mathfrak{R}_n$ denotes Rademacher complexity. where d_p is the effective parameter dimension of the hypothesis class.

The immunity bias term arises from SDL spectral collapse weighted by λ and stable-rank coefficient η and from curvature gating controlled by κ. Under smoothness and bounded-gradient assumptions a first-order analysis gives

$$\mathbb{E}\big[L_{\text{train}}(\hat{\theta}) - L_{\text{train}}^{\text{noSDL}}(\hat{\theta})\big] = O(\lambda\eta) + O(\kappa^2). \tag{42}$$

The reconstruction error from generators G_m contributes an approximation term ε_{rec}.

Combining the three contributions yields the final bound

$$\mathbb{E}_{\text{gen}} \le O\left(\sqrt{\frac{d_p}{n}}\right) + O(\lambda\eta) + O(\kappa^2) + \varepsilon_{\text{rec}}. \tag{43}$$

where ε_{rec} denotes generator approximation error.

The decomposition justifies adaptively shrinking λ, η, κ via bi-level hyper-gradient descent as training proceeds in order to reduce bias while preserving robustness.

A.8 Summary

The results above provide the requested Fréchet derivative expression, a deterministic certificate and a high-probability lower bound for the empirical covariance spectrum, explicit bias and variance amplification constants for the curvature-gated mask, a closed-form Neumann truncation certificate with doubling-rule guarantee, a regret bound for EXP3.P under sub-Gaussian information-gain losses, arguments for the Lipschitz spectral proxy and its practical computability for ViT encoders, and a Rademacher-complexity-based generalization decomposition that yields the stated combined bound. If numerical instantiation of the hidden constants is desired, substitute the problem-specific sub-Gaussian parameter σ_x, gradient norm bound G, variance trace V, and the Hessian spectral norm β into the above expressions to obtain fully explicit inequalities.

B Constants Instantiation

Table 12: Empirical instantiation of ModalImmune theoretical constants on CMU-MOSI.

Constant	Meaning	Empirical value (mean of 3 runs)
C	Bernstein matrix concentration constant	0.47
σ_x	sub-Gaussian parameter of embedding z	1.8
β	Hessian spectral norm $\|H\|_2$	6.1
α	inner learning rate	5×10^{-3}
G	gradient norm bound $\mathbb{E}\|g_0\|_2$	4.2
V	trace bound $\text{tr}(\Sigma_{g_0})$	190

The empirical procedure used to obtain the numbers in Table 12 is as follows. Ten thousand embeddings were sampled prior to training and the method-of-moments estimator described in [56] was used to fit the sub-Gaussian parameter,

from which the Bernstein constant C was derived. The Hessian spectral norm β was estimated by running Power Iteration at the start of each epoch on one thousand random mini-batches sampled from the training set and taking the maximum observed value across those trials. Gradient statistics used to produce G and V were collected over the first five thousand optimizer steps without gradient clipping; the reported values correspond to the empirical 95th percentiles of the observed gradient norms and trace estimates respectively. These procedures are deterministic given the random seeds and dataset partition and are intended to provide reproducible, practical instantiations of the abstract constants that appear in the theoretical certificates.

C Implementation Details

The online Neumann truncation tolerance is set relative to the current validation hyper-gradient norm. Specifically we use

$$\varepsilon \;=\; 0.02 \times \left\| \tfrac{\partial \mathcal{L}_{\mathrm{val}}}{\partial \theta} \right\|_2, \tag{44}$$

where $\|\cdot\|_2$ denotes the Euclidean norm of the validation gradient and the tolerance is recomputed once per epoch.

EXP3.P is configured with a conservative, theoretically-informed learning rate and a small exploration constant. The learning rate is set elementwise by

$$\eta_{\exp} \;=\; \min\Big\{0.5,\; \frac{\ln K}{KT}\Big\}, \tag{45}$$

where K denotes the number of arms (modalities) and T denotes the planned number of bandit rounds. The exploration parameter is fixed to $\gamma = 0.05$. Theoretical regret bounds are expressed per-sample and therefore the asymptotic guarantee does not depend on the employed mini-batch size; in practice small batches (for example 32) only affect empirical variance constants but do not change the established $O(\sqrt{TK \log K})$ scaling in expectation.

The curvature threshold used by curvature-gated masking is defined by

$$\tau \;=\; \frac{0.01}{\mathrm{lr}_{\mathrm{enc}}}, \tag{46}$$

where $\mathrm{lr}_{\mathrm{enc}}$ denotes the encoder learning rate at the current epoch and is taken as the instantaneous value produced by the cosine annealing schedule. This choice ties the curvature sensitivity to the current optimization scale and avoids manual retuning across different training phases.

Neumann inversion is initialized with a single-term approximation and doubled until the residual satisfies the tolerance in (44). Concretely, the initial truncation depth is

$$K_0 \;=\; 1, \tag{47}$$

and after each doubling the residual bound is re-evaluated; the truncation depth is capped at $K_{\max} = 10$ to bound computational cost. When the residual certificate falls below ε we accept the current K and use the resulting Neumann approximation for the hyper-gradient computation.

All hyperparameter defaults and the procedures above are used in the experiments unless stated otherwise in the main text. These explicit implementation choices remove ambiguity and allow the theoretical certificates reported in Section A to be instantiated numerically for reproducible evaluation.