UNDERSTANDING MISSING-MODALITY ROBUSTNESS THROUGH TARGET-RELEVANT REDUNDANCY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal models often suffer substantial performance degradation when partial modalities become unavailable. Existing approaches primarily focus on improving performance under incomplete inputs, while leaving unclear what information property makes multimodal model inherently robust to modality incompleteness. In this work, we study missing-modality robustness from the perspective of cross-modal redundancy. We formulate cross-modal redundancy as the target-relevant information shared between modalities, and establish its theoretical connection to robustness under modality incompleteness. Our analysis shows that larger cross-modal redundancy leads to a more favorable bound on the average prediction error when either modality is unavailable. Guided by this analysis, we develop a *Redundancy-Oriented Multimodal learning* (ROM) framework. Specifically, we introduce two complementary objectives that strengthen target-mediated cross-modal dependence while suppressing residual class-conditional dependence, thereby encouraging the model to obtain more target-relevant redundancy. Extensive experiments across different modality availability settings demonstrate improved robustness to missing modalities. Further analyses reveal a consistent correspondence between increased target-relevant redundancy and improved missing-modality performance, supporting our theoretical predictions. Together, our results establish target-relevant cross-modal redundancy as a principled information perspective for understanding and improving multimodal robustness.

1 INTRODUCTION

Multimodal learning aims to integrate heterogeneous observations from different modalities to support reliable prediction (Baltrušaitis et al., 2018). Most existing multimodal models, however, are developed under the assumption that all modalities are available during both training and inference. In real-world applications, this assumption can be easily violated due to sensor failures, privacy constraints, or acquisition costs. When one modality becomes unavailable, the model can only make predictions solely from the remaining observations, potentially losing information carried by the missing modality. Such modality incompleteness can substantially degrade model performance, making robustness to missing modalities a fundamental challenge for practical multimodal learning.

Existing approaches to modality incompleteness primarily improve robustness through input completion (Lin et al., 2021) or proxy objectives (Ma et al., 2022) under simulated missing modalities. Recent studies further suggest that representation characteristics, such as cross-modal alignment (Tian & Xu, 2021) and reduced modality dominance (Wang et al., 2024; Li et al., 2025), can affect robustness. However, the underlying property that makes multimodal representations inherently resilient to modality loss remains unclear. In particular, it remains unclear what information properties multimodal representations should capture to improve missing-modality robustness.

We address this question through the lens of **target-relevant cross-modal redundancy**, which measures the redundancy among modalities that is relevant to predicting the target. As shown in Figure 1 (left), attributes such as the bird's yellow crown and dark wings are reflected in both image and text and are informative for predicting the label, whereas other shared information is irrelevant to the target. Intuitively, when one modality becomes unavailable, such target-relevant redundant information can still be retained by the remaining modality. As shown in Figure 1 (right), we empirically observe that more redundancy indeed leads to less performance degradation on the missing modal-

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

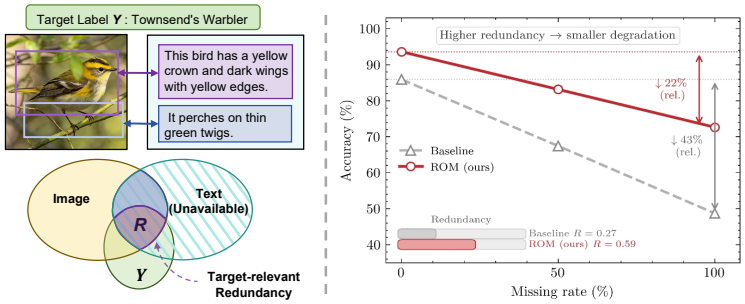


Figure 1: **Left**: Redundancy captures the target-relevant shared information between modalities. **Right**: Higher cross-modal redundancy leads to stronger robustness under missing modalities.

ity. Motivated by this observation, we define a data-level redundancy measure that characterizes the cross-modal dependence attributable to the modalities' shared relationship with the target. We further establish an information-theoretic connection between this redundancy and robustness to modality loss: **larger target-relevant cross-modal redundancy leads to a more favorable bound on the average prediction error when either modality is unavailable.** This result provides a principled explanation of why multimodal systems that preserve more target-relevant shared information can be more resilient to modality incompleteness.

Building on this theoretical explanation, we develop a redundancy-oriented learning framework that explicitly shapes multimodal representations toward greater target-relevant redundancy. We decompose redundancy learning into two complementary components and translate the information-theoretic objective into tractable training objectives. First, we construct label-consistent cross-modal interactions to strengthen target-mediated shared information. We then introduce a class-conditional decorrelation objective to suppress residual cross-modal dependence within each class, preventing arbitrary correlations from being treated as useful redundancy. Together, the two objectives selectively enhance target-relevant multimodal redundancy by strengthening desirable target-mediated dependence while removing undesirable residual dependence. Unlike approaches that primarily improve robustness through exposure to predefined missing-modality patterns, our framework directly shapes the information structure of multimodal representations. Extensive experiments demonstrate consistent improvements under modality incompleteness, while further analyses reveal a clear association between increased cross-modal redundancy and improved missing-modality robustness, consistent with our theoretical analysis.

Our contributions are summarized as follows:

- We provide a new information perspective on missing-modality robustness by identifying target-relevant cross-modal redundancy as a key underlying property.
- We establish a theoretical connection between target-relevant cross-modal redundancy and prediction robustness under modality loss.
- Guided by this analysis, we develop a redundancy-oriented representation learning framework and empirically validate across different modality availability settings.

2 Related Work

2.1 Learning with Incomplete Modalities

Existing approaches to modality incompleteness primarily aim to improve predictive performance when certain inputs are unavailable. One line of work reconstructs missing modalities from the observed ones, thereby approximating the unavailable information before prediction (Pham et al., 2019; Sun et al., 2024; Ke et al., 2025; Hai et al., 2026). Another common strategy simulates modality incompleteness during training by masking or dropping modalities, explicitly adapting the model to different missing patterns (Neverova et al., 2016; Maheshwari et al., 2024). More recent approaches further adapt model components to different missing patterns or reduce excessive dependence on dominant modalities (Wang et al., 2024; Li et al., 2025). At the representation level, representation alignment, modality-invariant learning, and cross-modal knowledge compensation have also been explored to improve robustness to missing modalities (Lin & Hu, 2023; Saeed et al., 2024; Zhao et al., 2025). Existing methods mainly improve robustness to missing modalities through

empirical designs, with limited study of why these methods work. Recent work has started to explore possible reasons, such as modality dominance and modality invariance (Wang et al., 2024; Saeed et al., 2024). However, what kind of information multimodal representations should learn to remain robust to missing modalities is still unclear.

2.2 Redundancy in Multimodal Learning

Redundancy characterizes information shared across modalities and has long played an important role in multimodal representation learning. Early cross-modal learning methods implicitly exploit such shared information by aligning representations of paired modalities, allowing one modality to provide supervision for another (Radford et al., 2021; Tschannen et al., 2025). However, generic cross-modal dependence does not necessarily reflect information that is useful for a downstream task. This distinction has motivated information-theoretic formulations of task-relevant redundancy (Liang et al., 2023; Dufumier et al., 2025). For example, the multi-view information bottleneck (Federici et al., 2020) defines two views as redundant with respect to a target when one view provides no additional target information given the other, and uses their shared information to learn sufficient representations. Recent studies have further shown that target-relevant redundancy can improve multimodal robustness (Wörtwein et al., 2024; Nguyen et al., 2025). When the same predictive information is available from multiple modalities, the model can still rely on this shared evidence when one modality is noisy or unavailable. However, existing studies mainly use redundancy as a learning signal or an empirical property, while its connection to robustness remains less understood. Our work studies this connection from an information-theoretic perspective. We explicitly define a measure of target-relevant redundancy and establish its theoretical relationship with robustness to modality incompleteness.

3 Theoretical Analysis

3.1 Preliminaries

Throughout this section, uppercase letters denote random variables and lowercase letters their realizations. Let X_1 and X_2 denote two input modalities and Y the target label, jointly distributed according to the underlying data distribution $p(X_1, X_2, Y)$. A multimodal model f_θ maps the inputs into modality-specific representations Z_1 and Z_2, which are subsequently used to predict Y. We write $H(\cdot)$ for entropy and $I(\cdot\,;\cdot)$ for mutual information.

3.2 Target-Relevant Redundancy Measure

We define the *target-relevant cross-modal redundancy indicator* at the data level as

$$R_X = I(X_1; X_2) - I(X_1; X_2 \mid Y). \tag{1}$$

The first term, $I(X_1; X_2)$, captures the overall statistical dependence between the two input modalities, while the second term, $I(X_1; X_2 \mid Y)$, measures the residual dependence that remains after conditioning on the prediction target. Their difference therefore characterizes the portion of cross-modal dependence associated with the common relationship of the two modalities to Y.

Importantly, R_X is determined by the underlying data distribution and is independent of the learned model. It thus characterizes an intrinsic task-dependent information structure: different multimodal tasks may contain different degrees of target-relevant cross-modal redundancy even before any representation learning is performed.

3.3 From Redundancy to Missing-Modality Robustness

We next analyze how this intrinsic target-relevant redundancy relates to robustness when one modality becomes unavailable. Let $P(e_1)$ and $P(e_2)$ denote the prediction error probabilities when X_1 and X_2 are missing, respectively. We derive the following bound:

$$\frac{\Phi_X - \log 2}{\log(|\mathcal{Y}| - 1)} \le \frac{P(e_1) + P(e_2)}{2} \le 1 - \exp(-\Phi_X), \qquad \text{where} \quad \Phi_X = H(Y) - \frac{R_X + I(X_1, X_2; Y)}{2}. \tag{2}$$

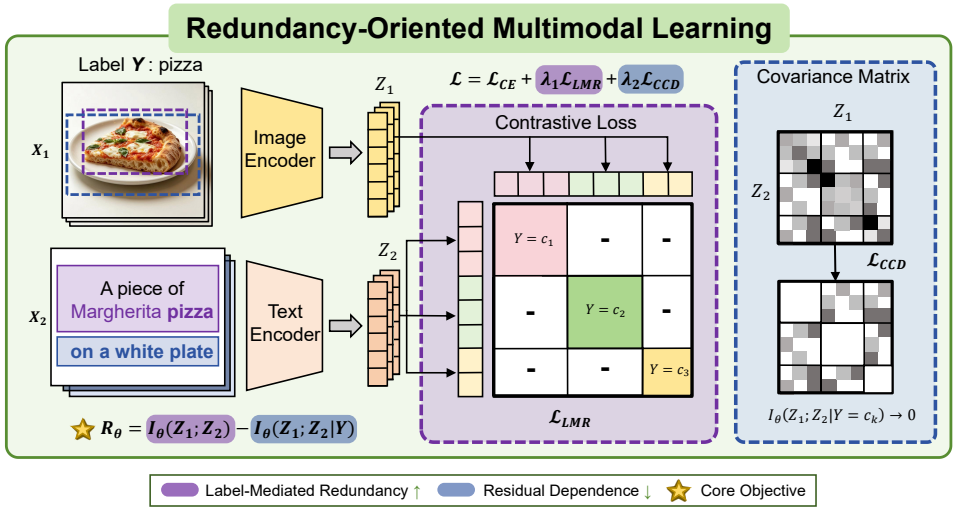


Figure 2: Framework of Redundancy-Oriented Multimodal learning (ROM).

Here, $I(X_1, X_2; Y)$ measures the total predictive information jointly provided by the two modalities, whereas R_X characterizes how much of the target-relevant information structure is shared across them. For a fixed amount of joint predictive information $I(X_1, X_2; Y)$, a larger R_X leads to a smaller Φ_X, and consequently a more favorable bound on the average prediction error under modality loss. Proofs and assumptions are deferred to the appendix.

This result identifies target-relevant cross-modal redundancy as an intrinsic information property associated with missing-modality robustness. Since this intrinsic structure is determined by the data distribution, learning cannot alter R_X; instead, a robust multimodal model should preserve and effectively capture such target-relevant redundancy in its learned representations.

4 METHOD

The analysis in Section 3 identifies target-relevant cross-modal redundancy as an intrinsic information property underlying robustness to missing modalities. Although such data-level redundancy is determined by the underlying task distribution, representation learning determines the extent to which this target-relevant shared information is preserved in the learned representations. Motivated by this distinction, we define the representation-level redundancy as

$$R_\theta = I_\theta(Z_1; Z_2) - I_\theta(Z_1; Z_2 \mid Y). \tag{3}$$

We focus on learning representations with larger R_θ while retaining sufficient predictive information about the target for better robustness.

Direct optimization of these information-theoretic quantities is generally intractable. As illustrated in Figure 2, we introduce two complementary structural objectives to indirectly optimize target-relevant cross-modal redundancy. The first promotes cross-modal dependence, encouraging the two modalities to consistently capture shared target semantics. The second reduces residual cross-modal dependence within class, encouraging the learned representations to capture category-level shared structure rather than spurious correlations among samples and modalities from the same class. Together, these objectives provide a theory-guided mechanism for shaping the learned representations toward greater target-relevant cross-modal redundancy.

4.1 LABEL-MEDIATED REDUNDANCY LEARNING

A representation structure with larger R_θ should exhibit strong cross-modal dependence that is relevant to the prediction target, while avoiding dependence arising from task-irrelevant correlations. Simply increasing $I_\theta(Z_1; Z_2)$, however, does not distinguish between these two sources of dependence. We therefore seek to promote cross-modal dependence that can be attributed to the shared relationship of the two modalities with the target variable Y.

To characterize this desired statistical structure, we construct the surrogate distribution

$$q_\theta(z_1, z_2, y) = p(y)p_\theta(z_1 \mid y)p_\theta(z_2 \mid y), \tag{4}$$

which satisfies $Z_1 \perp Z_2 \mid Y$ under q_θ. Consequently, any marginal dependence between Z_1 and Z_2 under q_θ can only arise through their shared dependence on Y, rather than through residual instance-specific correlations after conditioning on the target. We therefore use $I_{q_\theta}(Z_1; Z_2)$ as a surrogate objective for promoting label-mediated cross-modal dependence.

This construction admits a direct operational interpretation. For a fixed label $Y = y$, the conditional distribution under q_θ is obtained by independently drawing

$$z_1 \sim p_\theta(z_1 \mid y), \qquad z_2 \sim p_\theta(z_2 \mid y). \tag{5}$$

Accordingly, all cross-modal representations sharing the same label can be treated as positive pairs, without requiring instance-level correspondence. In a mini-batch, considering all label-consistent cross-modal combinations empirically approximates sampling from the conditional product distribution $p_\theta(z_1 \mid y)p_\theta(z_2 \mid y)$. This construction therefore emphasizes target-mediated dependence while removing the requirement that positive pairs originate from the same multimodal instance.

Directly estimating $I_{q_\theta}(Z_1; Z_2)$ is generally intractable. We therefore adopt a contrastive surrogate motivated by the standard density-ratio interpretation of contrastive learning. We denote the representation of the i-th sample from modality m by $z_{m,i}$ and ℓ_2-normalize all representations. The scaled similarity $z_{1,i}^\top z_{2,j}/\tau$, where τ is a learnable temperature, serves as a contrastive critic for distinguishing label-consistent cross-modal pairs from mismatched ones. This provides a tractable mechanism for promoting the label-mediated dependence characterized by q_θ. For each class c represented in the mini-batch, we define the set of sample indices belonging to that class as $\mathcal{S}_c = \{i : y_i = c\}$. All cross-modal combinations within $\mathcal{S}_c$ are treated as label-consistent positive pairs. The directional objective from modality 1 to modality 2 is

$$\mathcal{L}_{1\to 2} = -\frac{1}{N}\sum_c \sum_{i\in\mathcal{S}_c} \frac{1}{|\mathcal{S}_c|} \sum_{j\in\mathcal{S}_c} \log \frac{\exp(z_{1,i}^\top z_{2,j}/\tau)}{\sum_{k=1}^{N} \exp(z_{1,i}^\top z_{2,k}/\tau)}. \tag{6}$$

Since target-relevant redundancy is inherently symmetric between the two modalities, we apply the same objective in the reverse direction and define the Label-Mediated Redundancy loss as

$$\mathcal{L}_{\mathrm{LMR}} = \frac{1}{2}\left(\mathcal{L}_{1\to 2} + \mathcal{L}_{2\to 1}\right). \tag{7}$$

The resulting objective takes the form of multi-positive cross-modal contrastive learning, but its role here is to promote the label-mediated dependence captured by q_θ, rather than to directly estimate the representation-level redundancy indicator R_θ. Accordingly, $\mathcal{L}_{\mathrm{LMR}}$ encourages cross-modal representations associated with the same target semantics to exhibit stronger dependence, while relaxing the requirement of strict instance-level correspondence.

4.2 Class-Conditional Decorrelation

While $\mathcal{L}_{\mathrm{LMR}}$ promotes target-mediated cross-modal dependence, the conditional term in R_θ captures the residual dependence between the two modality representations after the target label is given. We therefore seek to reduce $I_\theta(Z_1; Z_2 \mid Y)$. Direct estimation of conditional mutual information is generally intractable. We therefore characterize class-conditional dependence through the normalized cross-correlation matrix. For a fixed class $Y = c$, let $\Gamma_c = \Sigma_{11,c}^{-1/2}\Sigma_{12,c}\Sigma_{22,c}^{-1/2}$. Under a class-conditional Gaussian assumption and sufficiently weak cross-modal correlations,

$$I_\theta(Z_1; Z_2 \mid Y = c) \approx \frac{1}{2}\|\Gamma_c\|_F^2. \tag{8}$$

This approximation motivates suppressing second-order cross-modal dependence within each class. In practice, we use class-conditional cross-covariance as a simple mini-batch surrogate for this dependence, avoiding the need to estimate full class-conditional covariance structure. The detailed derivation is provided in the Appendix A.1.

Specifically, we first ℓ_2-normalize the representations from both modalities to control feature scale. For each class c in a mini-batch, let $Z_1^{(c)}, Z_2^{(c)} \in \mathbb{R}^{N_c \times D}$ denote the corresponding normalized

representations, which are then centered within each class as

$$\widetilde{Z}_m^{(c)} = Z_m^{(c)} - \frac{1}{N_c}\mathbf{1}\mathbf{1}^\top Z_m^{(c)}, \qquad m \in \{1, 2\}, \tag{9}$$

and compute the empirical class-conditional cross-covariance matrix

$$C_c = \frac{1}{N_c - 1}\left(\widetilde{Z}_1^{(c)}\right)^\top \widetilde{Z}_2^{(c)}. \tag{10}$$

Penalizing the cross-covariance matrix, we define the Class-Conditional Decorrelation (CCD) loss:

$$\mathcal{L}_{\mathrm{CCD}} = \frac{1}{|\mathcal{C}_B|}\sum_{c \in \mathcal{C}_B} \|C_c\|_F^2, \tag{11}$$

where $\mathcal{C}_B$ denotes the set of classes used to compute the loss. $\mathcal{L}_{\mathrm{CCD}}$ can be interpreted as a practical second-order proxy for conditional cross-modal dependence.

4.3 Overall Objective

The overall training objective of our method can be denoted as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_{\mathrm{LMR}}\mathcal{L}_{\mathrm{LMR}} + \lambda_{\mathrm{CCD}}\mathcal{L}_{\mathrm{CCD}}, \tag{12}$$

where λ_{LMR} and λ_{CCD} control the strengths of the two regularization terms. $\mathcal{L}_{\mathrm{LMR}}$ promotes cross-modal dependence mediated by the target variable, whereas $\mathcal{L}_{\mathrm{CCD}}$ suppresses residual dependence after conditioning on the target. By jointly optimizing these auxiliary objectives alongside the main task loss, the model learns to complete the multimodal prediction task while simultaneously building a representation structure with stronger target-mediated dependence and weaker target-conditioned residual dependence, consistent with greater target-relevant cross-modal redundancy.

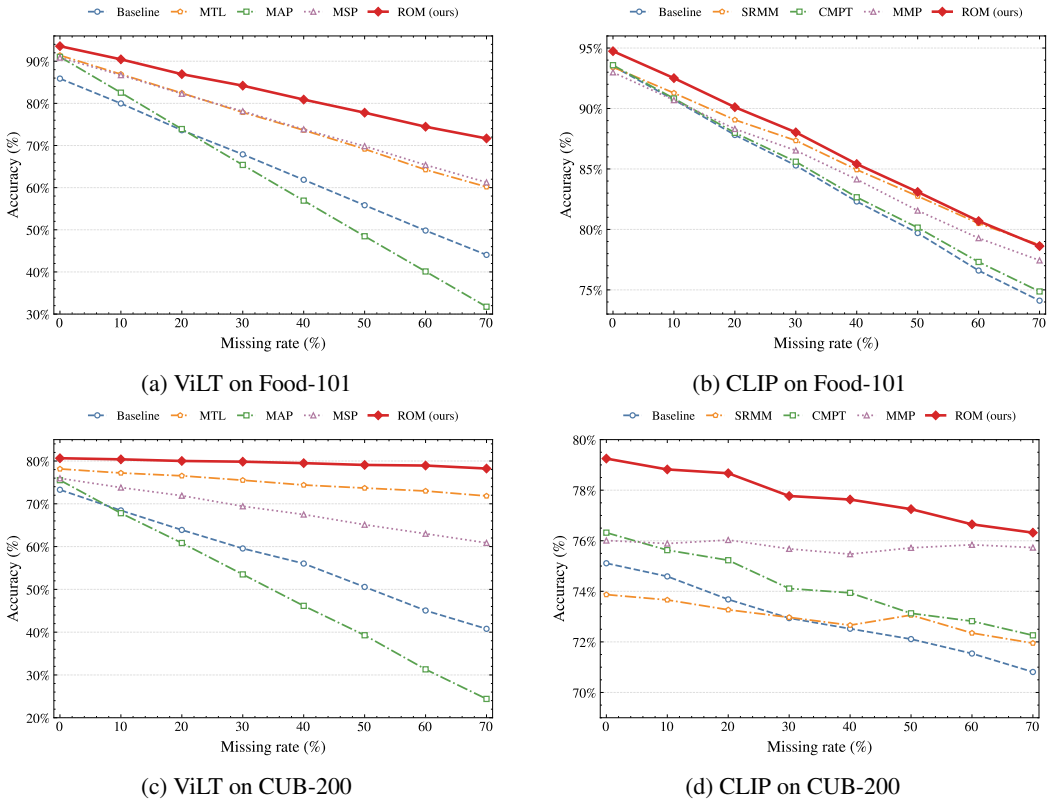
5 Experiment

5.1 Experimental Setting

Datasets We evaluate our method on three text-image multimodal benchmarks, FOOD-101 (Bossard et al., 2014), CUB-200 (Wah et al., 2011), and HatefulMemes (Kiela et al., 2020), covering diverse prediction tasks. FOOD-101 is a food classification dataset with 101 categories, while CUB-200 is a fine-grained bird classification dataset containing 200 species. HatefulMemes is a more challenging binary classification benchmark for detecting hateful content from paired images and text; the corresponding experiments are reported in the Appendix B.

Comparison To evaluate whether our method generalizes across different multimodal fusion architectures, we conduct experiments under both single-stream and dual-stream settings. For the single-stream setting, we adopt ViLT (Kim et al., 2021) as the common backbone, where image and text inputs are jointly processed within a unified transformer, and compare against Baseline, MTL (Ma et al., 2022), MAP (Lee et al., 2023), and MSP (Jang et al., 2024). For the dual-stream setting, we adopt CLIP to independently extract image and text representations, which are subsequently integrated through late fusion. We implement SRMM (Saeed et al., 2024), CMPT (Reza et al., 2025), and MMP (Nezakati et al., 2025) under the same CLIP-based dual-stream framework for comparison. All methods within each setting use the same backbone and follow the same missing-modality protocol to ensure a fair comparison.

Experimental Setting. For ViLT-based methods, we finetune the ViLT-B/32 pretrained checkpoint using AdamW with a learning rate of 1×10^{-2}, a weight decay of 0.02, and a batch size of 512 for 20 epochs. We apply linear warmup over the first 10% of training steps, followed by cosine annealing. All images are resized to 384×384, and training is performed with FP16 mixed precision on two GPUs. For CLIP-based methods, we use CLIP ViT-B/32 as the backbone, with a learning rate of 1×10^{-2} for the prediction head and 1×10^{-5} for the backbone. All other training settings remain the same as ViLT. For ROM, we set $\lambda_{\mathrm{LMR}} = 0.5$ and $\lambda_{\mathrm{CCD}} = 1$. To quantify redundancy, we adopt the Redundancy Bottleneck (Kolchinsky, 2024), a PID-based measure of redundant information. We use it as a computable proxy for target-relevant information shared across modalities and examine its correlation with missing-modality robustness.



(a) ViLT on Food-101

(b) CLIP on Food-101

(c) ViLT on CUB-200

(d) CLIP on CUB-200

Figure 3: Robustness evaluation when models are trained with complete multimodal inputs. The results show the performance degradation under different missing ratios during inference.

Incompleteness setting We consider two practical missing-modality scenarios, depending on whether modality incompleteness occurs only at inference time or during both training and inference. For certain incompleteness scenario, we use predefined missing rate p, following which each modality is randomly removed. Note that each sample should have at least one modality available. We evaluate models over multiple missing ratios to simulate different levels of modality availability.

5.2 Main Result

Training with Full Modality. We first consider a challenging setting where models are trained exclusively with complete multimodal inputs but encounter missing modalities during inference. This setting evaluates whether models can implicitly learn representations that are robust to modality absence without observing incomplete samples during training. We evaluate models under different missing ratios, with the results shown in Figure 3. We observe that existing approaches trained with complete multimodal inputs often suffer from substantial performance degradation as modality availability decreases. Notably, dual-stream architectures generally exhibit stronger robustness than single-stream counterparts, suggesting that maintaining relatively independent modality-specific representations can alleviate the sensitivity to missing inputs. Nevertheless, training on complete multimodal data alone does not guarantee robust multimodal representations, as models can still rely heavily on modality-specific information rather than learning sufficient cross-modal redundancy. In contrast, our method consistently achieves stronger robustness across different missing ratios and architectures. These results demonstrate that explicitly enhancing target-relevant cross-modal redundancy encourages models to retain predictive information that is shared across modalities, thereby improving their ability to maintain performance under modality incompleteness.

Training with Incompleteness We further evaluate a more practical setting where modality incompleteness is already present during training. In this setting, models must learn useful multimodal representations from partially observed training examples while remaining robust to missing modalities at inference time. The missing ratio can be different between training and inference, making incompleteness robustness a hard task. As shown in Table 1, we make three observations.

Table 1: Accuracy (%) under different modality availability settings on Food-101 and CUB-200. The best and second-best results are highlighted in **bold** and underline, respectively.

Dataset	Training		Inference		Single-Stream					Dual-Stream				
	Image	Text	Image	Text	Baseline	MTL	MAP	MSP	ROM	Baseline	SRMM	CMPT	MMP	ROM
Food-101	100%	30%	100%	30%	59.72	67.02	70.67	70.53	**77.47**	78.09	78.31	76.70	76.23	**80.14**
			30%	100%	65.83	**86.44**	32.12	58.14	85.22	86.05	86.59	87.36	86.29	**87.59**
			65%	65%	62.71	76.81	51.52	64.51	**81.50**	82.15	82.18	82.24	81.09	**83.49**
	30%	100%	100%	30%	34.34	38.67	29.11	42.69	**63.37**	72.29	**74.89**	72.94	73.42	73.64
			30%	100%	82.67	86.89	86.39	85.95	**88.14**	88.15	88.20	87.42	87.15	**88.73**
			65%	65%	58.48	62.84	57.73	64.24	**75.80**	80.30	**81.36**	80.06	80.29	81.15
	65%	65%	100%	30%	57.09	60.74	67.94	68.12	**75.36**	77.83	77.47	74.62	74.37	**78.84**
			30%	100%	80.79	87.00	85.85	85.99	**87.81**	87.48	87.55	87.37	87.22	**88.43**
			65%	65%	68.75	73.89	77.00	77.14	**81.69**	82.81	82.04	80.98	80.80	**83.39**
CUB-200	100%	30%	100%	30%	69.12	65.33	72.97	73.35	**79.53**	73.63	72.09	74.09	69.38	**77.77**
			30%	100%	21.68	**24.78**	21.73	22.01	23.85	32.38	44.11	**46.72**	38.89	36.18
			65%	65%	46.31	45.20	47.46	48.31	**52.30**	53.24	57.63	**61.05**	54.85	56.90
	30%	100%	100%	30%	19.95	27.80	19.18	25.35	**36.66**	56.82	60.37	62.24	62.46	**63.29**
			30%	100%	43.42	50.41	48.24	48.48	**56.40**	49.12	49.10	55.13	54.73	**55.63**
			65%	65%	32.02	39.37	34.12	37.31	**46.62**	53.40	54.44	**59.20**	58.63	58.99
	65%	65%	100%	30%	60.32	72.14	67.47	67.92	**75.30**	69.31	69.21	69.99	68.45	**72.66**
			30%	100%	39.85	**53.24**	46.84	46.62	52.95	46.84	48.76	48.29	51.10	**53.05**
			65%	65%	50.09	62.06	58.42	57.30	**64.64**	59.06	58.63	59.84	59.39	**63.12**

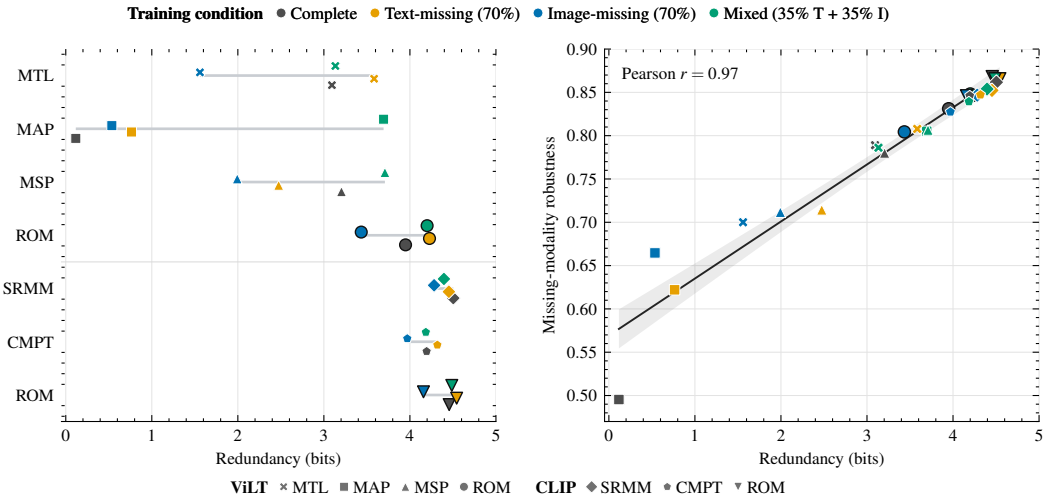


Figure 4: **Left:** Redundancy comparison over different methods and training setting on Food-101. **Right:** Correlation between target-relevant redundancy and missing-modality robustness.

First, existing models are sensitive to the discrepancy between training and inference modality availability. When the missing pattern at inference differs from that during training, their performance often degrades substantially, indicating limited generalization to unseen incomplete scenarios. Second, existing missing-modality approaches can improve robustness under incomplete training settings, especially when the training and inference missing distributions are aligned. However, their improvements become less consistent under distribution shifts, suggesting that they mainly adapt to specific missing patterns. Notably, MSP and SRMM remain relatively effective despite such train–test condition mismatches, indicating stronger robustness across different missing-modality conditions. Third, our method consistently achieves competitive or superior performance across both single-stream and two-stream architectures. Notably, the improvements are more prominent under mismatched training and inference missing ratios, demonstrating that our method learns more generalizable and robust multimodal representations.

5.3 Redundancy Verification

Target-relevant redundancy provides a consistent representation-level indicator of robustness to modality incompleteness. As shown in Figure 4 (left), different robustness methods learn substantially different levels of cross-modal redundancy. On ViLT, conventional methods such as MAP

Table 2: Ablation of the proposed objectives on target-relevant redundancy bottleneck (RB) and robustness under modality incompleteness on ViLT backbone.

Dataset	Method	RB ↑	ΔRB	Text ↓70%	Image ↓70%	Both ↓70%	Avg. ↑	ΔAvg.
Food-101	Baseline	1.825	–	44.07	75.90	60.05	60.01	–
	$\mathcal{L}_{\mathrm{CCD}}$	3.156	+1.331	60.79	84.90	72.81	72.83	+12.82
	$\mathcal{L}_{\mathrm{LMR}}$	4.011	+2.186	70.84	86.22	78.51	78.52	+18.51
	ROM	**4.075**	**+2.250**	**71.65**	**86.38**	**79.03**	**79.02**	**+19.01**
CUB-200	Baseline	0.224	–	40.90	22.35	31.93	31.73	–
	$\mathcal{L}_{\mathrm{CCD}}$	0.242	+0.018	31.08	28.74	30.27	30.03	-1.70
	$\mathcal{L}_{\mathrm{LMR}}$	2.121	+1.897	67.62	28.91	48.15	48.23	+16.50
	ROM	**3.378**	**+3.154**	**78.24**	**38.16**	**58.28**	**58.23**	**+26.50**

and MSP exhibit relatively low or condition-sensitive redundancy, whereas ROM consistently maintains a high redundancy level across different training incompleteness settings. A similar pattern is observed on the CLIP-based architecture, where robust methods remain concentrated in the high-redundancy regime. These results suggest that robustness is associated not only with the amount of shared target-relevant information, but also with its stability under changes in the training missing pattern. Figure 4 (right) further shows that this representation-level property closely corresponds to downstream robustness. Across architectures, methods, and training conditions, redundancy is strongly correlated with accuracy under modality incompleteness ($r = 0.97$). Together, these observations provide empirical evidence that models preserving more target-relevant information across modalities tend to remain more reliable when one modality becomes unavailable.

5.4 Ablation study

We further validate our claim that ROM improves robustness to modality incompleteness by increasing target-relevant redundancy. To this end, we ablate the two proposed objectives respectively and examine the contribution of each component to both redundancy and missing-modality robustness. The results are reported in Table 2.

On Food-101, both objectives substantially increase target-relevant redundancy and improve robustness under modality incompleteness. Combining the two objectives further achieves the highest redundancy and overall robustness, suggesting that the two objectives can provide complementary benefits when both modalities contain sufficient target-relevant information. A different pattern emerges on CUB-200. While $\mathcal{L}_{\mathrm{LMR}}$ remains highly effective, $\mathcal{L}_{\mathrm{CCD}}$ alone yields only a marginal increase in redundancy and slightly decreases average robustness. This observation is consistent with the role of $\mathcal{L}_{\mathrm{CCD}}$: reducing conditional cross-modal dependence is most effective when both modalities already contain sufficient target-relevant information, i.e., when both $I(Z_1; Y)$ and $I(Z_2; Y)$ are sufficiently large. When one modality carries limited target-relevant information, enforcing conditional independence cannot compensate for the missing information and may instead over-regularize useful cross-modal dependence.

Overall, these results suggest that $\mathcal{L}_{\mathrm{LMR}}$ serves as the primary mechanism for increasing target-relevant redundancy and improving missing-modality robustness, while $\mathcal{L}_{\mathrm{CCD}}$ provides a complementary effect when sufficient target information is available in both modalities. Both objective collaboratively enhance the redundancy and thus improve the missing-modality robustness.

6 Conclusion

In this work, we study missing-modality robustness from the perspective of target-relevant cross-modal redundancy. We establish an information-theoretic connection between redundancy and prediction robustness under modality loss, showing that greater target-relevant redundancy is associated with more favorable error bounds. Based on this insight, we propose ROM, a redundancy-oriented multimodal learning framework that strengthens target-mediated cross-modal dependence while suppressing task-irrelevant conditional dependence. Experiments across different architectures and modality availability settings consistently demonstrate improved robustness, while further analyses confirm a strong correspondence between learned redundancy and missing-modality performance. Overall, our work moves beyond empirical adaptation to missing modalities and provides a principled information perspective for understanding and improving robust multimodal learning.

AI DISCLOSURE STATEMENT

In this work, we used generative AI tools to assist with checking mathematical derivations and proofs, creating and refining figures for clearer presentation, suggesting color schemes for scientific visualization, and identifying relevant literature to improve the coverage of related work. We carefully reviewed and verified all AI-assisted outputs, including each step of the mathematical derivations, the relevance and correctness of the identified literature, and the accuracy with which the generated or refined figures represent the intended scientific content. We take full responsibility for the final content of this work, including all text, claims, proofs, figures, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model architectures, training hyperparameters, optimization settings, and missing-modality protocols used in our experiments. The datasets, preprocessing procedures, evaluation metrics, and benchmark-specific settings are also documented in the main paper and supplementary material. We additionally provide complete derivations and assumptions for the theoretical results in the appendix. The source code, training configurations, and evaluation scripts will be released upon acceptance to facilitate reproduction of the reported results.

REFERENCES

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443, 2018.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Benoit Dufumier, Javiera Castillo Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? In *International Conference on Learning Representations*, volume 2025, pp. 5408–5432, 2025.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. *arXiv preprint arXiv:2002.07017*, 2020.

Yan Hai, Shanqi Lu, Zhizhong Liu, Ling Shang, Hongxiang Sun, and Jing Wang. Sequential translation-based multimodal sentiment analysis under uncertain missing modalities. *Scientific Reports*, 16:21018, 2026.

Jaehyuk Jang, Yooseung Wang, and Changick Kim. Towards robust multimodal prompting with missing modalities. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8070–8074. IEEE, 2024.

Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25864–25873, 2025.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Artemy Kolchinsky. Partial information decomposition: Redundancy as information bottleneck. *Entropy*, 26(7):546, 2024.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14943–14952. IEEE, 2023.

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

Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36:27351–27393, 2023.

Ronghao Lin and Haifeng Hu. Missmodal: Increasing robustness to missing modality in multimodal sentiment analysis. *Transactions of the Association for Computational Linguistics*, 11:1686–1702, 2023.

Yijie Lin, Yuanbiao Gou, Zitao Liu, Boyun Li, Jiancheng Lv, and Xi Peng. Completer: Incomplete multi-view clustering via contrastive prediction. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11169–11178. IEEE, 2021.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18156–18165. IEEE, 2022.

Harsh Maheshwari, Yen-Cheng Liu, and Zsolt Kira. Missing modality robustness in semi-supervised multi-modal semantic segmentation. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1009–1019. IEEE, 2024.

Natalia Neverova, Christian Wolf, Graham W. Taylor, and Florian Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, 2016.

Niki Nezakati, Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M Salman Asif. Mmp: Towards robust multi-modal learning with masked modality projection. In *2025 IEEE International Conference on Big Data (BigData)*, pp. 1480–1485. IEEE, 2025.

Duy A Nguyen, Abhi Kamboj, and Minh N Do. Robult: Leveraging redundancy and modality specific features for robust multimodal learning. *arXiv preprint arXiv:2509.03477*, 2025.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 6892–6899, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Md Kaykobad Reza, Ameya Patil, Mashhour Solh, and M Salman Asif. Robust multimodal learning via cross-modal proxy tokens. *arXiv preprint arXiv:2501.17823*, 2025.

Muhammad Saad Saeed, Shah Nawaz, Muhammad Zaigham Zaheer, Muhammad Haris Khan, Karthik Nandakumar, Muhammad Haroon Yousaf, Hassan Sajjad, Tom De Schepper, and Markus Schedl. Modality invariant multimodal learning to handle missing modalities: A single-branch approach. *arXiv preprint arXiv:2408.07445*, 2024.

Yuhang Sun, Zhizhong Liu, Quan Z. Sheng, Dianhui Chu, Jian Yu, and Hongxiang Sun. Similar modality completion-based multimodal sentiment analysis under uncertain missing modalities. *Information Fusion*, 110:102454, 2024.

Yapeng Tian and Chenliang Xu. Can audio-visual integration strengthen robustness under multimodal attacks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5601–5611, 2021.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Hao Wang, Shengda Luo, Guosheng Hu, and Jianguo Zhang. Gradient-guided modality decoupling for missing-modality robustness. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(14):15483–15491, 2024.

Torsten Wörtwein, Nicholas B Allen, Jeffrey F Cohn, and Louis-Philippe Morency. Smurf: Statistical modality uniqueness and redundancy factorization. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pp. 339–349, 2024.

Yihan Zhao, Wei Xi, Xiao Fu, and Jizhong Zhao. Enhancing multimodal model robustness under missing modalities via memory-driven prompt learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 2458–2466, 2025.

A Formulation and Proof

A.1 Details about Class-Conditional Decorrelation loss

We seek to reduce

$$I_\theta(Z_1; Z_2 \mid Y).$$

Direct estimation of conditional mutual information is generally intractable. We instead derive a second-order surrogate based on class-conditional cross-modal dependence. For a fixed class $Y = c$, we assume the joint representation is approximately Gaussian,

$$\begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \Bigg| Y = c \sim \mathcal{N}\left(\begin{bmatrix} \mu_{1,c} \\ \mu_{2,c} \end{bmatrix}, \begin{bmatrix} \Sigma_{11,c} & \Sigma_{12,c} \\ \Sigma_{21,c} & \Sigma_{22,c} \end{bmatrix} \right).$$

Let

$$R_c = \Sigma_{11,c}^{-1/2} \Sigma_{12,c} \Sigma_{22,c}^{-1/2}$$

denote the normalized cross-correlation matrix. Under the Gaussian assumption,

$$I_\theta(Z_1; Z_2 \mid Y = c) = -\frac{1}{2} \log\det\left(I - R_c R_c^\top\right).$$

For sufficiently weak cross-modal correlations,

$$-\log\det(I - A) \approx \mathrm{Tr}(A),$$

which gives

$$I_\theta(Z_1; Z_2 \mid Y = c) \approx \frac{1}{2} \|R_c\|_F^2.$$

If the within-class marginal representations are additionally approximately whitened,

$$\Sigma_{11,c} \approx I, \qquad \Sigma_{22,c} \approx I,$$

then $R_c \approx \Sigma_{12,c}$, yielding

$$I_\theta(Z_1; Z_2 \mid Y = c) \approx \frac{1}{2} \|\Sigma_{12,c}\|_F^2.$$

This motivates suppressing within-class cross-covariance as a practical second-order surrogate for reducing conditional cross-modal dependence.

We implement this surrogate at the mini-batch level. To control feature scale, we first ℓ_2-normalize the representations from each modality. For each class c in the mini-batch, let

$$Z_1^{(c)}, Z_2^{(c)} \in \mathbb{R}^{N_c \times D}$$

denote the corresponding normalized representations. We center them along the sample dimension,

$$\widetilde{Z}_m^{(c)} = Z_m^{(c)} - \frac{1}{N_c} \mathbf{1}\mathbf{1}^\top Z_m^{(c)}, \qquad m \in \{1, 2\},$$

and compute the empirical class-conditional cross-covariance

$$C_c = \frac{1}{N_c - 1} \left(\widetilde{Z}_1^{(c)}\right)^\top \widetilde{Z}_2^{(c)}.$$

We then define the Class-Conditional Decorrelation loss as

$$\mathcal{L}_{\mathrm{CCD}} = \frac{1}{|\mathcal{C}_B|} \sum_{c \in \mathcal{C}_B} \|C_c\|_F^2,$$

where $\mathcal{C}_B$ denotes the set of classes used to compute the loss. In practice, we compute the decorrelation objective within each class and average uniformly across classes, preventing frequent classes from dominating the optimization. We penalize the entire cross-modal covariance matrix to reduce residual second-order dependence between the two modalities after conditioning on the target. Importantly, zero cross-covariance does not imply statistical independence in general. Therefore, $\mathcal{L}_{\mathrm{CCD}}$ is not a direct estimator of $I_\theta(Z_1; Z_2 \mid Y)$, but a tractable second-order surrogate for reducing conditional cross-modal dependence under the assumptions discussed above. Together with $\mathcal{L}_{\mathrm{LMR}}$, it provides a practical mechanism for increasing target-relevant cross-modal redundancy by strengthening shared target information while suppressing residual within-class dependence.

A.2 Proof for Equation 2

Let $P(e_1)$ and $P(e_2)$ denote the prediction error probabilities when X_1 and X_2 are missing, respectively. We derive the following bound:

$$\frac{\Phi_X - \log 2}{\log(|\mathcal{Y}| - 1)} \leq \frac{P(e_1) + P(e_2)}{2} \leq 1 - \exp(-\Phi_X),$$
$$\text{where} \qquad \Phi_X = H(Y) - \frac{R_X + I(X_1, X_2; Y)}{2}.$$

We denote X as samples and Y as ground truth. A trained model make prediction $\tilde{Y} = g(X)$

We define an indicator random variable E to denote whether the estimate $\tilde{Y} = g(X)$ is incorrect:

$$E = \begin{cases} 1 & \text{if } \tilde{Y} \neq Y, \\ 0 & \text{if } \tilde{Y} = Y. \end{cases}$$

Now, consider the joint entropy $H(E, Y \mid \tilde{Y})$. By the chain rule of entropy, this can be decomposed in two ways:

$$\begin{aligned} H(E, Y \mid \tilde{Y}) &= H(Y \mid \tilde{Y}) + H(E \mid Y, \tilde{Y}) \\ &= H(E \mid \tilde{Y}) + H(Y \mid E, \tilde{Y}). \end{aligned}$$

Notice that once Y and $\tilde{Y}$ are given, E is completely determined, so $H(E \mid Y, \tilde{Y}) = 0$. Equating the two expressions yields:

$$H(Y \mid \tilde{Y}) = H(E \mid \tilde{Y}) + H(Y \mid E, \tilde{Y}).$$

Let us analyze $H(Y \mid E, \tilde{Y})$. By the law of total probability:

$$\begin{aligned} H(Y \mid E, \tilde{Y}) &= H(Y \mid E = 0, \tilde{Y}) \cdot P(E = 0) \\ &\quad + H(Y \mid E = 1, \tilde{Y}) \cdot P(E = 1) \\ &= 0 \cdot P(E = 0) + H(Y \mid E = 1, \tilde{Y}) \cdot P(e) \\ &= H(Y \mid E = 1, \tilde{Y}) \cdot P(e), \end{aligned}$$

where $P(e) = P(E = 1)$ is the probability of error in our estimate. The first term vanishes since if $E = 0$, then $\tilde{Y} = Y$, and conditioned on $\tilde{Y}$, Y is known exactly and thus entropy is zero.

On the other hand, if $E = 1$, then $\tilde{Y} \neq Y$. Comprehending the upper equation, we can have:

$$H(Y \mid \tilde{Y}) = H(E \mid \tilde{Y}) + H(Y \mid E = 1, \tilde{Y}) P(e),$$

A.2.1 Lower Bound Derivation

Given $\tilde{Y}$, Y must take values in the remaining $|\mathcal{Y}| - 1$ possibilities, so

$$H(Y \mid E = 1, \tilde{Y}) \leq \log(|\mathcal{Y}| - 1).$$

Therefore,

$$H(Y \mid E, \tilde{Y}) \leq \log(|\mathcal{Y}| - 1) \cdot P(e).$$

Now, consider $H(E \mid \tilde{Y})$. By the property that conditioning reduces entropy,

$$H(E \mid \tilde{Y}) \leq H(E).$$

Recall that E is a Bernoulli random variable with probability of error $e = P(E = 1)$. Thus,

$$H(E) = H_b(e),$$

where $H_b(e)$ denotes the binary entropy function. Hence,

$$H(E \mid \tilde{Y}) \leq H_b(e) \leq \log 2.$$

Combining these results, we obtain:

$$H(Y \mid \tilde{Y}) \leq P(e) \log(|\mathcal{Y}| - 1) + \log 2.$$

Lastly, since $Y \to X \to \tilde{Y}$ forms a Markov chain, by the data processing inequality,

$$I(Y;\tilde{Y}) \le I(Y;X).$$

This implies that

$$H(Y \mid \tilde{Y}) \ge H(Y \mid X).$$

So we conclude

$$P(e) \ge \frac{H(Y|X) - \log 2}{\log(|\mathcal{Y}| - 1)}.$$

The inequality above holds for an arbitrary decoder $\tilde{Y} = g(X)$. When X_1 is missing, predictions are formed from the remaining modality X_2, so

$$P(e_1) \ge \frac{H(Y \mid X_2) - \log 2}{\log(|\mathcal{Y}| - 1)}.$$

Symmetrically,

$$P(e_2) \ge \frac{H(Y \mid X_1) - \log 2}{\log(|\mathcal{Y}| - 1)}.$$

Averaging the two inequalities yields

$$\frac{P(e_1) + P(e_2)}{2} \ge \frac{\Phi_X - \log 2}{\log(|\mathcal{Y}| - 1)}.$$

where

$$\Phi_X := \frac{H(Y \mid X_1) + H(Y \mid X_2)}{2} = H(Y) - \frac{I(X_1;Y) + I(X_2;Y)}{2}.$$

By the chain-rule identity

$$R_X = I(X_1;X_2) - I(X_1;X_2 \mid Y) = I(X_1;Y) + I(X_2;Y) - I(X_1,X_2;Y),$$

we obtain

$$\Phi_X = H(Y) - \frac{R_X + I(X_1,X_2;Y)}{2},$$

The bound requires $|\mathcal{Y}| \ge 3$; the binary case $\log(|\mathcal{Y}| - 1) = 0$ is excluded and can be handled separately.

A.2.2 Upper Bound Derivation

To establish the upper bound of the error rate, we assume maximum a posteriori (MAP) decoding, i.e., $\hat{Y}(x) = \arg\max_y P(y \mid x)$. Let $p_{\max}(x) = \max_y P(y \mid x)$ denote the posterior probability of the MAP prediction. For any posterior, the Shannon entropy is lower-bounded by the negative log of its largest atom:

$$H(Y \mid X = x) \ge -\log p_{\max}(x),$$

with equality if and only if the posterior is a point mass. Taking the expectation over X yields:

$$H(Y \mid X) \ge \mathbb{E}\big[-\log p_{\max}(X)\big].$$

Since $-\log(\cdot)$ is convex, Jensen's inequality gives:

$$\mathbb{E}\big[-\log p_{\max}(X)\big] \ge -\log \mathbb{E}\big[p_{\max}(X)\big].$$

Under MAP decoding, the probability of a correct prediction is $\mathbb{E}[p_{\max}(X)] = 1 - P(e)$, and therefore:

$$H(Y \mid X) \ge -\log\big(1 - P(e)\big).$$

Here and below, $\log$ denotes the natural logarithm. Exponentiating both sides:

$$\exp(-H(Y \mid X)) \le 1 - P(e) \implies P(e) \le 1 - \exp(-H(Y \mid X))$$

Considering the two modalities X_1 and X_2 with error rates $p(e_1)$ and $p(e_2)$, we average the entropy bounds:

$$\frac{H(Y \mid X_1) + H(Y \mid X_2)}{2} \ge \frac{-\log(1 - p(e_1)) - \log(1 - p(e_2))}{2}$$

Let $\bar{P}(e) = \frac{p(e_1)+p(e_2)}{2}$. Since $f(x) = -\log(1-x)$ is a convex function, by Jensen's inequality:

$$\frac{-\log(1 - p(e_1)) - \log(1 - p(e_2))}{2} \ge -\log(1 - \bar{P}(e))$$

Let Φ denote the average conditional entropy $\frac{H(Y|X_1)+H(Y|X_2)}{2}$. We conclude:

$$\Phi \ge -\log(1 - \bar{P}(e)) \implies \bar{P}(e) \le 1 - \exp(-\Phi)$$

Table 3: Accuracy (%) under different modality availability settings on HatefulMemes dataset. The best and second-best results are highlighted in **bold** and underline, respectively.

Training		Inference		Dual-Stream				
Image	**Text**	**Image**	**Text**	**Baseline**	**SRMM**	**CMPT**	**MMP**	**ROM**
100%	100%	100%	30%	65.87	66.72	65.34	67.55	**67.70**
		30%	100%	66.49	64.27	**68.06**	65.33	66.58
		65%	65%	**64.33**	63.66	62.45	63.91	64.11
100%	30%	100%	30%	**68.83**	64.43	63.02	59.30	68.57
		30%	100%	63.48	63.71	65.68	60.08	**66.59**
		65%	65%	63.86	59.89	61.08	56.75	**64.76**
30%	100%	100%	30%	64.85	63.06	60.92	61.82	**66.43**
		30%	100%	**65.53**	58.89	65.37	59.91	61.07
		65%	65%	62.34	59.61	60.45	57.96	**63.11**
65%	65%	100%	30%	64.40	63.73	59.14	63.81	**65.60**
		30%	100%	64.85	60.26	62.23	59.47	**66.60**
		65%	65%	**63.49**	63.38	57.52	59.83	62.53

B Robustness on Low-Redundancy Data

We further evaluate robustness on HatefulMemes, a more challenging dataset with substantially lower cross-modal redundancy. In this setting, learning from incomplete modalities becomes considerably harder, since the missing information is less likely to be recoverable from the remaining modality. As shown in Table 3, existing missing-modality methods therefore provide less consistent gains and can even underperform the baseline under some training–inference configurations. In contrast, ROM remains among the strongest methods across most settings, particularly under mismatched modality availability. These results suggest that, even when naturally shared information is limited, ROM can better exploit the available target-relevant redundancy, leading to more robust representations under modality incompleteness.