Distributionally Robust Multimodal Machine Learning

Peilin Yang
Univeristy of Cambridge
py245@cam.ac.uk

Yu Ma
University of Wisconsin, Madison
yu.ma@wisc.edu

Abstract

We consider the problem of robust multimodal machine learning. Existing approaches often rely on merging modalities on the feature level (early fusion) or heuristic uncertainty modeling, which downplays modality-aware effects and provide limited insights. We propose a novel distributionally robust optimization (DRO) framework that aims to study both the theoretical and practical insights of multimodal machine learning. We first justify this setup and show the significance of this problem through complexity analysis. We then establish both generalization upper bounds and minimax lower bounds which provide performance guarantees. These results are further extended in settings where we consider encoder-specific error propogations. Empirically, we demonstrate that our approach improves robustness in both simulation settings and real-world datasets. Together, these findings provide a principled foundation for employing multimodal machine learning models in high-stakes applications where distributional uncertainty is unavoidable.

1 Introduction

Multimodal learning is increasingly incorporated ins machine learning modeling frameworks to build systems in high-stake environments such as healthcare and policy making Soenksen et al. [2022], Lewington et al. [2024]. However, deploying these models remain challenging despite their performance advantages, with one main criticism on such system's volatility when data change drastically across input settings.

Multimodal learning under uncertainty raises some naturally important practical and theoretical questions. Previous works Bezirganyan et al. [2024] have incorporated varying belief distributions over different modalities with prior evidence of specific modality's reliability and provides a principled approach to prevent overfinance that result in wrong predictions when modalities disagree. Works have also been done to extend previous unimodal neural processes to multimodal NP settings Jung et al. [2023], and have been applied to meta-learning settings to incorporate such uncertainty-aware framework Almecija et al. [2022]. However, recent approaches have largely focused on the demonstration of incorporating uncertainties improve out-of-distribution performance. It remains to be justified why such setting is fundamentally different from merging all features across modalities and modeling the uncertainties of each feature. Another line of principled approach is distributionally robust optimization (DRO), a well-established framework that has been applied to capture uncertainty in data covariate shifts across various machine learning models Liu et al. [2025], Lewington et al. [2024], Blanchet et al. [2019], Zhu et al. [2022], Cho et al. [2024], Shafieezadeh-Abadeh et al. [2015], Chen and Paschalidis [2018], Bertsimas et al. [2019]. In this paper, we introduce a novel DRO multimodal machine learning and provide theoretical and practical insights not previously studied.

The main contribution of this paper is to introduce a new framework for robust multimodal machine learning using distributionally robust optimization. We provide theoretical justifications for the significance of the modality-aware formulation and demonstrate its fundamental difference from early fusion through complexity analysis. This construction enables us to obtain closed-form, tractable formulations of the upper and lower bounds of the risk, which translates to performance guarantees for users in practice. These bounds also incorporate the consideration of correlation structures across modalities. Finally, we provide empirical results in both simulation and real-world datasets that demonstrate our approach improves computational out of sample performance.

Preprint.

The rest of the paper is organized as follows. Section 2 outlines the problem setting of the DRO formulation and provides justification for the modality-aware setup. Section 3 presents the generalization upper bound and risk lower bounds. Section 4 demonstrates the effectiveness of the DRO framework in both simulation and real-world settings. We conclude the paper in Section 5. In the Appendix, we provide all the proofs as well as other additional computational results.

2 A Framework for Robust Multimodal Learning

2.1 Problem Setup and Formulation

We consider a multimodal machine learning prediction problem with K modalities. Let $X = (X^{(1)}, \dots, X^{(K)})$ denote the input, Y the target, and N the sample size. We consider predictors of the form $f \circ g$, where $g = (g^{(1)}, \dots, g^{(K)})$ maps inputs to one-dimensional embeddings $Z^{(k)} = g^{(k)}(X^{(k)}) \in \mathbb{R}$, and f aggregates the embeddings $\{Z^{(k)}\}$ to predict Y. Specifically, adopting a one-dimensional embedding $Z^{(k)}$ isolates the role of cross-modal correlation without loss of conceptual generality and provides us with analytical results.

Assumption 1 (Copula Dependence)**.** *Let P_X and Q_X be the nominal and perturbed joint distributions over X, respectively, with marginals $P_X^{(k)}$ and $Q_X^{(k)}$. We analyze robustness to covariate shift in X (and thus Z), and assume that $P_{Y|X}$ is fixed. By Sklar's theorem, $P_X(x^{(1)}, \dots, x^{(K)}) = C(P_X^{(1)}(x^{(1)}), \dots, P_X^{(K)}(x^{(K)}))$, for a copula C that captures cross-modal dependence.*
This decomposition separates each modality's marginal behavior from their dependence structure.

Assumption 2 (Shared copula)**.** *P_X and Q_X share the same copula C. Let c be the copula density, the joint measures decompose as $dP_X = c\Big(P_X^{(1)}, \dots, P_X^{(K)}\Big)\, dP_X^{(1)} \cdots dP_X^{(K)}$, and $dQ_X = c\Big(Q_X^{(1)}, \dots, Q_X^{(K)}\Big)\, dQ_X^{(1)} \cdots dQ_X^{(K)}$. where modality-wise constraints $D_{\chi^2}(Q_X^{(k)} \parallel P_X^{(k)}) \le \rho_k$.*

The shared copula assumption ensures that our following robustness analysis focuses on shifts in the marginals while preserving cross-modal dependence. This is inline with realistic data scenarios where marginal distributions such as image quality could shift, but modality correlations remain relatively stable. We also acknowledge that choice of chi-square is due to its tractable formulation, distance metrics such as Wasserstein distance should be further analyzed as well.

2.2 Why Modality-Aware DRO?

A natural question is why we need to introduce this framework if one could simply concatenate all features at the data stage and apply a traditional DRO formulation? In other words, what is the significance of treating modalities separately, and how does our setup differ from this canonical alternative? Following conventional machine learning terminologies, we refer the first approach as early fusion, and the latter as late fusion. Based on the problem setting introduced above, among the K modalities, we have for each of the modalities dimensions $\{d_i\}_{i=1}^K$, and covariates $X^{(i)} = \{X^{(ij)}\}_{j=1}^{d_i}$. Let $D = \sum_i d_i$. We define the two approaches as follow:

- *Early fusion*: Concatenate all features and learn $h : \mathbb{R}^D \to Y$ directly.
- *Late fusion*: Learn 1-D embedding functions $f_i : X^{(i)} \to Z^{(i)} \in \mathbb{R}$, then learn separate prediction function $g : \{Z^{(k)}\} \in \mathbb{R}^K \to Y$.

We first observe that these two approaches exhibit different computational complexity behaviors under different algorithmic structures.

Proposition 1. *Using linear structures and ordinary least squares (OLS), early fusion has complexity $O\big(ND^2 + D^3\big)$, and late fusion has complexity $O\big(N \sum_i d_i^2 + \sum_i d_i^3 + NK^2 + K^3\big)$.*

This indicates that the modality-wise approach is typically more computationally efficient when total number of modalities K is large and individual embedding dimensions d_i are small.

Proposition 2. *Under gradient-based training (e.g., SGD), early and late fusion requires $\tilde{O}(ND)$ and $\tilde{O}\big(ND + NK\big)$ per epoch, respectively. If parallel processing is available, the wall-clock time per sample can be reduced from $\tilde{O}(D)$ to $\tilde{O}(\max_i\{d_i\})$.*
We observe that instead of using an unconstrained worst-case divergence in the concatenated feature space, the modality-aware formulation provides a bound that is decomposed across modalities and explicitly incorporates cross-modality correlations.

3 Correlation-Aware Worst-Case Risk

In this section, we derive the generalization gap of the DRO formulation, and derive how this gap tightens under assumptions of the encoder function. We conclude with the risk lower bound obtained using a two-point distribution construction.

We adopt the following lemma to ensure that our downstream arguments of the distributional shift on the original input raw data remain valid in the embedding space.

Lemma 1 (Data processing Inequality). *If $Z^{(k)} = g^{(k)}(X^{(k)})$, then $D_{\chi^2}\left(Q_Z^{(k)} \parallel P_Z^{(k)}\right) \leq D_{\chi^2}\left(Q_X^{(k)} \parallel P_X^{(k)}\right) \leq \rho_k$. Under Assumption 2, the joint embedded divergence admits a correlation-aware expansion. Let $\frac{dQ_Z^{(k)}}{dP_Z^{(k)}} = 1 + \epsilon_k, \mathbb{E}_{P_Z^{(k)}}[\epsilon_k] = 0, \mathbb{E}_{P_Z^{(k)}}[\epsilon_k^2] \leq \rho_k$. Define γ_{ij} as copula-induced correlations in embedding space so that $\mathbb{E}[\epsilon_i \epsilon_j] \leq \gamma_{ij}\sqrt{\rho_i \rho_j}$ for $\gamma_{ij} \geq 0$ (and the inequality reverses if $\gamma_{ij} < 0$). Neglecting higher-order terms,*

$$D_{\chi^2}(Q_Z \parallel P_Z) = \mathbb{E}_{P_Z}\left[\left(\prod_{k=1}^K (1 + \epsilon_k) - 1\right)^2\right] \leq \underbrace{\sum_{k=1}^K \rho_k + 2\sum_{i<j} |\gamma_{ij}| \sqrt{\rho_i \rho_j}}_{:= \mathcal{B}} . \tag{1}$$

Lemma 2 (Risk decomposition). *Given data $\mathcal{D} = (X, Y)$, a bounded loss function $\ell(\cdot) \in [0, M_\ell]$ and hypothesis classes $\mathcal{F}, \mathcal{G}$ where $f \in \mathcal{F}$ and $g \in \mathcal{G}$. For a fixed $f \circ g$, define the worst-case risk with a χ^2-ambiguity set with radius as the correlation-aware $\mathcal{B}$ in* (1). *If $Var(l) < \infty$, it has the closed form: $r(f \circ g, P_X) := \sup_{Q_X \colon D_{\chi^2}(Q_Z \| P_Z) \leq \mathcal{B}} \mathbb{E}_{P_{Y|X} \times Q_X}\left[\ell(f \circ g(X), Y)\right] = \mathbb{E}[\ell] + \sqrt{\mathcal{B}}\,\sqrt{\mathrm{Var}(\ell)}$.*

Lemma 3 (Lipschitz property of sample SD). *Let $f(X_1, \ldots, X_N) = \left(\frac{1}{N}\sum_{i=1}^N (X_i - \bar{X})^2\right)^{1/2}$ with $\bar{X} = \frac{1}{N}\sum_{i=1}^N X_i$. Then f is $N^{-1/2}$-Lipschitz w.r.t. the ℓ_2-norm.*

From Lemma 2 and 3, we derive the following bounds.

Lemma 4 (Finite-sample deviation). *Let $\hat{P}_X$ denote the empirical distribution over $\{X_i\}_{i=1}^N$. with probability at least $1 - 2e^{-t}$, for an absolute constant C: $\left|r(f \circ g, \hat{P}_X) - \mathbb{E}_{P_X}\left[r(f \circ g, \hat{P}_X)\right]\right| \leq \sqrt{\mathcal{B}}\,\sqrt{\frac{2t}{N}}\,M_\ell + \sqrt{\frac{t}{2N}}\,M_\ell$, and $\left|\mathbb{E}_{P_X}[r(f \circ g, \hat{P}_X)] - r(f \circ g, P_X)\right| \leq \sqrt{\mathcal{B}}\,\frac{C}{N}$.*

Combining the two inqualities from Lemma 4:

Theorem 1 (Generalization upper bound). *Let $\ell \in [0, M_\ell]$ and assume Assumption 2. Then, with probability at least $1 - 2e^{-t}$,*

$$\left|r(f \circ g, \hat{P}_X) - r(f \circ g, P_X)\right| \;\leq\; \sqrt{\mathcal{B}}\,\sqrt{\frac{2t}{N}}\,M_\ell + \sqrt{\frac{t}{2N}}\,M_\ell + \sqrt{\mathcal{B}}\,\frac{C}{N},$$

where $\mathcal{B}$ is the correlation-aware ambiguity radius in (1).

The upper bound can be considered as a general certificate of confidence that in essence translates to a guarantee. Consider a high-stake environment, such as healthcare, where risk prediction (i.e., heart failure in the next 24-hours) is often the primary goal of most machine learning models. Such upper bound provides care providers an estimate that "given a certain level of uncertainty, the model's performance degrade will not exceed X". This is especially important to provide in deployment as we can quantify the amount of performance and ad-hoc prepare users to adapt in advance.

We can further tighten this bound under assumptions of how perturbations in the original input propagate through the encoder as follows.

Theorem 2 (Encoder–robust upper bound). *Assume further that $\ell(\cdot)$ is L_ℓ–Lipschitz, and g is modality–wise Lipschitz with constant $L_{g,i}$ in modality i. Consider an encoder perturbation only in modality i, writing $\Delta_i(x^{(i)}) := \hat{f}^{(i)}(x^{(i)}) - f^{(i)}(x^{(i)})$. For any $t > 0$, with probability at least $1 - 2e^{-t}$ over a sample of size n that defines $\widehat{P}_X$*

$$\begin{aligned} \left|\, r(g \circ \hat{f}, \widehat{P}_X) - r(g \circ f, P_X) \,\right| \;&\leq L_\ell L_{g,i}\Big(\mathbb{E}_{\widehat{P}_X}\big[|\Delta_i(X^{(i)})|\big] + \sqrt{\mathcal{B}}\,\sqrt{\mathrm{Var}_{\widehat{P}_X}\big(|\Delta_i(X^{(i)})|\big)}\Big) \\ &\quad + \sqrt{\mathcal{B}}\Big(\sqrt{\tfrac{2t}{n}}\,M_\ell + \tfrac{C}{n}\Big) + \sqrt{\tfrac{t}{2n}}\,M_\ell \,. \end{aligned}$$

Intuitively, this indicates that the worst-case risk cannot blow up faster than the size of the change of an encoder on a given modality—even when the data distribution shifts. In practice, this allows practitioners to safely swap, compress, or quantize an encoder (or run with a partially missing/degraded modality) with knowledge of a hard ceiling on the performance degrade.

Beyond the generalization gap, it is also important to characterize the intrinsic difficulty of estimating the worst-case risk. Consider again a healthcare setting, the lower bound can provide a realistic estimate to the care providers by setting expectations on what is the best a certain model can do under the distributional shift. We do so using the following minimax quantity:

$$\mathfrak{M}_N := \inf_{\hat{f}\circ\hat{g}} \sup_{P_X \in \mathcal{P}} \mathbb{E}\big|r(\hat{f}\circ\hat{g}, \hat{P}_X) - r(f\circ g, P_X)\big|.$$

Theorem 3 (Risk lower bound)**.** *If the loss function ℓ is bounded by $L > 0$, then given large $M > 0$ and sufficiently large sample size N, $\mathfrak{M}_N \;\geq\; \frac{L}{4N\log(M-1)}$.*

This provides a conservative baseline and does not preclude stronger lower bounds when $\mathcal{P}$ excludes degenerate distributions or when modality-specific divergences (and their correlations) grow with K.

4 Numerical experiment

We conduct both simulation and real-world computational experiments to illustrate the advantage of our DRO formulation. Additional experiments can be found in Appendix C. All experiments were run on a CPU-only laptop with an Intel Core i7-12700H (14 cores/20 threads) and 16 GB RAM. We used Python 3.12. All experiments are completed within 21-36 minutes.

4.1 Simulation Setup and Results

Let us consider four modalities $\{m^{(k)} \in \mathbb{R}^5\}_{k=1}^4$. Under the *training* distribution P, draw $m^{(1)} \sim \mathcal{N}(0\cdot\mathbf{1}_5,\ I_5)$ and $m^{(k)} = w\,m^{(1)} + (1-w)\,\varepsilon^{(k)}$ for $k = 2,3,4$ with $w = 0.7$ and $\varepsilon^{(k)} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2 I_5)$, $\sigma_\varepsilon = 0.05$. Define the response $Y = \mathbf{1}_5^\top \sum_{k=1}^4 m^{(k)} \;+\; 10\,\mathbf{1}\{\bar{m}^{(1)} > 1\} \;-\; 10\,\mathbf{1}\{\bar{m}^{(1)} < -1\}$, $\quad \bar{m}^{(1)} := \frac{1}{5}\mathbf{1}_5^\top m^{(1)}$. The *test* distribution Q differs only by a mean shift in the primary modality, $m^{(1)} \sim \mathcal{N}\big(1.5\cdot\mathbf{1}_5,\ I_5\big)$, while the conditional construction of $m^{(2)}, m^{(3)}, m^{(4)}$ is unchanged, so the shift propagates across modalities through w. We observe that by sacrificing limited performance on the majority, it improves robustness on the minority subgroup.

ρ	**Whole (MSE)**	**Whole (DRO)**	**Minor (MSE)**	**Minor (DRO)**
0.10	4.990 ± 0.163	4.230 ± 0.133	5.521 ± 0.186	4.603 ± 0.150
0.50	4.651 ± 0.148	3.708 ± 0.105	5.103 ± 0.167	3.984 ± 0.119
1.20	5.124 ± 0.177	3.825 ± 0.101	5.676 ± 0.200	4.134 ± 0.116
2.00	5.060 ± 0.169	3.570 ± 0.086	5.577 ± 0.190	3.822 ± 0.097

Table 1: Whole vs. minority subgroup for MSE and χ^2-DRO across optimization radius ρ.

4.2 Real-world Case Studies

We consider two real-world settings: journalism and healthcare, and provide detailed descriptions of each of these dataset in Appendix C. The training and testing sets are stratified chronologically. For each experiment, we train a 2-layer MLP with Adam learning rate of 0.005. We compare the standard non-robust approach to our DRO approach with $\rho = 0.5$.

Dataset	**Method**	**Median Acc.**	**IQR [Q1, Q3]**	**Std. Dev.**
Journalism	Canonical	0.5798	[0.5666, 0.5924]	0.0180
Journalism	DRO	0.5853	[0.5743, 0.5951]	0.0164
Healthcare	Canonical	0.6027	[0.5903, 0.6144]	0.0185
Healthcare	DRO	0.6148	[0.6026, 0.6266]	0.0169

Table 2: Test accuracy over three real world cases. Acc. = Accuracy.

5 Conclusion

In this paper, we introduce a novel distributionally robust framework for multimodal learning that explicitly accounts for modality-specific shifts and cross-modal correlations. We established the significance of this problem through theoretical complexity analysis, and provide guarantees for its generalization. Empirical results across simulations and real-world datasets confirmed that our formulation improves prediction performance. Future work will aim to extend to incorporate more complex correlation structures and tighter analytical bounds.

References

Cesar Almecija, Apoorva Sharma, and Navid Azizan. Uncertainty-aware meta-learning for multimodal task distributions, 2022. URL `https://arxiv.org/abs/2210.01881`.

Dimitris Bertsimas, Jack Dunn, Colin Pawlowski, and Ying Daisy Zhuo. Robust classification. *INFORMS Journal on Optimization*, 1(1):2–34, January 2019. ISSN 2575-1492. doi: 10.1287/ijoo.2018.0001. URL `http://dx.doi.org/10.1287/ijoo.2018.0001`.

Grigor Bezirganyan, Sana Sellami, Laure Berti-Équille, and Sébastien Fournier. Multimodal learning with uncertainty quantification based on discounted belief fusion. 2024. doi: 10.48550/ARXIV.2412.18024. URL `https://arxiv.org/abs/2412.18024`.

Jose Blanchet, Yang Kang, and Karthyek Murthy. Robust wasserstein profile inference and applications to machine learning. *Journal of Applied Probability*, 56(3):830–857, September 2019. ISSN 1475-6072. doi: 10.1017/jpr.2019.49. URL `http://dx.doi.org/10.1017/jpr.2019.49`.

Ruidi Chen and Ioannis Ch. Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *Journal of Machine Learning Research*, 19(13):1–48, 2018. URL `http://jmlr.org/papers/v19/17-295.html`.

Wooseong Cho, Taehyun Hwang, Joongkyu Lee, and Min-hwan Oh. Randomized exploration for reinforcement learning with multinomial logistic function approximation, 2024. URL `https://arxiv.org/abs/2405.20165`.

Myong Chol Jung, He Zhao, Joanna Dipnall, and Lan Du. Beyond unimodal: Generalising neural processes for multimodal uncertainty estimation, 2023. URL `https://arxiv.org/abs/2304.01518`.

Aiden Lewington, Alekhya Vittalam, Anshumaan Singh, Anuja Uppuluri, Arjun Ashok, Ashrith Mandayam Athmaram, Austin Milt, Benjamin Smith, Charlie Weinberger, Chatanya Sarin, Christoph Bergmeir, Cliff Chang, Daivik Patel, Daniel Li, David Bell, Defu Cao, Donghwa Shin, Edward Kang, Edwin Zhang, Enhui Li, Felix Chen, Gabe Smithline, Haipeng Chen, Henry Gasztowtt, Hoon Shin, Jiayun Zhang, Joshua Gray, Khai Hern Low, Kishan Patel, Lauren Hannah Cooke, Marco Burstein, Maya Kalapatapu, Mitali Mittal, Raymond Chen, Rosie Zhao, Sameen Majid, Samya Potlapalli, Shang Wang, Shrenik Patel, Shuheng Li, Siva Komaragiri, Song Lu, Sorawit Siangjaeo, Sunghoo Jung, Tianyu Zhang, Valery Mao, Vikram Krishnakumar, Vincent Zhu, Wesley Kam, Xingzhe Li, and Yumeng Liu. Creating a cooperative ai policymaking platform through open source collaboration, 2024. URL `https://arxiv.org/abs/2412.06936`.

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.

Jiashuo Liu, Tianyu Wang, Henry Lam, Hongseok Namkoong, and Jose Blanchet. Dro: A python library for distributionally robust optimization in machine learning. *arXiv preprint arXiv:2505.23565*, 2025.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Soroosh Shafieezadeh-Abadeh, Peyman Mohajerin Esfahani, and Daniel Kuhn. Distributionally robust logistic regression, 2015. URL `https://arxiv.org/abs/1509.09259`.

Luis R Soenksen, Yu Ma, Cynthia Zeng, Leonard Boussioux, Kimberly Villalobos Carballo, Liangyuan Na, Holly M Wiberg, Michael L Li, Ignacio Fuentes, and Dimitris Bertsimas. Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ digital medicine*, 5(1):149, 2022.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Shixiang Zhu, Liyan Xie, Minghe Zhang, Rui Gao, and Yao Xie. Distributionally robust weighted k-nearest neighbors. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 29088–29100. Curran Associates, Inc., 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/file/bb0d37c6210f84abfa0fd7709edb30cb-Paper-Conference.pdf`.

A Qualitative Advantage of Modality-Aware Fusion

Beyond computational advantages, several other important benefits motivate our modality-wise approach:

1. **Specialized encoders:** Each encoder f_i can be tailored to its specific modality (e.g., CNNs for images, Transformers for text, MLPs for tabular data). This typically produces superior feature extraction compared to a single monolithic architecture attempting to handle all data types. The f_i encoders can be pretrained on large unimodal datasets or leverage off-the-shelf models (ResNet, BERT, etc.), requiring only the training (or light fine-tuning) of g. This approach substantially improves sample efficiency and represents a key reason why late fusion aligns more closely with the essence of multimodal learning.
2. **Robustness to missing modalities:** When modality i is missing for certain data points, one can simply skip f_i. In contrast, early fusion approaches typically require imputation strategies or retrained models. This represents a significant practical advantage in real-world multimodal systems.
3. **Handling heterogeneous data characteristics:** Different modalities may have varying frame rates (e.g., video vs. audio) or spatial dimensions. Per-modality encoders enable principled downsampling and aggregation before fusion. This natural variation across modalities motivates our consideration of covariate shift at the modality level, as each modality possesses its own inherent characteristics. Generalization capabilities vary across modalities, making Distributionally Robust Optimization (DRO) an ideal framework for enhancing overall model generalizability.
4. **Information bottleneck regularization:** The embedding process creates an information bottleneck Tishby et al. [2000], which theoretically reduces overfitting by compressing irrelevant intra-modality noise while preserving task-relevant signals. This perspective remains somewhat controversial, particularly given observed anomalies where additional modalities sometimes degrade performance. Several scholars (e.g., Liang et al. [2023]) have attempted to explain these phenomena through information-theoretic frameworks.

These considerations demonstrate the necessity of addressing covariate shift at the modality level, with DRO playing a crucial role in enhancing model generalization capabilities.

B Technical Proofs

Proof. (Proof of Proposition 1) Let N be the number of samples, K the number of modalities, and d_i the (embedding) dimension of modality i. Set $D := \sum_{i=1}^{K} d_i$.

- **Early fusion (single OLS on concatenated features).** Form the $N \times D$ design matrix $X = [X^{(1)} \ \cdots \ X^{(K)}]$ and solve $\min_w \|Xw - y\|_2^2$. Using normal equations or QR/Cholesky gives the standard costs: (i) form $X^\top X$ in $O(ND^2)$ flops and $X^\top y$ in $O(ND)$; (ii) factorize/solve the $D \times D$ system in $O(D^3)$ (solve back-substitution is $O(D^2)$ and is dominated). Hence the overall complexity is $O(ND^2 + D^3)$.
- **Late (modality-wise) fusion.**
  - Stage 1: For each modality i, fit an OLS on the $N \times d_i$ block $X^{(i)}$, costing $O(Nd_i^2 + d_i^3)$ by the same argument as above. Summing over i gives $\sum_{i=1}^{K} O(Nd_i^2 + d_i^3)$.
  - Stage 2: Fuse the K per-modality outputs (one scalar per modality, or a K-vector) via a linear head. This is an OLS in dimension K with cost $O(NK^2 + K^3)$.
  - Combining both stages yields

$$O\Big(\underbrace{\sum_{i=1}^{K}(Nd_i^2 + d_i^3)}_{\text{per-modality fits}} \ + \ \underbrace{NK^2 + K^3}_{\text{fusion head}}\Big).$$

This proves the stated bounds. □

Proof. (Proof of Proposition 2) We count one *epoch* as a full pass over N samples with first-order (stochastic) gradients. All $\tilde{O}(\cdot)$ bounds hide constant and polylogarithmic factors, and assume dense features; for sparse data, replace D by the average number of nonzeros per sample.

- **Early fusion.** With a linear (or shallow) model over the concatenated D-dimensional input, the forward/backward cost per sample is $\tilde{O}(D)$, so one epoch costs $\tilde{O}(ND)$.
- **Late (modality-wise) fusion.** Per sample, the gradient passes through each modality-specific block $X^{(i)}$ (cost $\tilde{O}(d_i)$) and then through the K-dimensional fusion head (cost $\tilde{O}(K)$). Summing over modalities gives per-sample cost $\tilde{O}\big(\sum_i d_i + K\big) = \tilde{O}(D+K)$, hence per epoch
$$\tilde{O}\big(N(D+K)\big) = \tilde{O}(ND + NK).$$
- **Parallelism across modalities.** If K modality blocks are processed in parallel (e.g., separate devices/streams) and gradients are synchronized only at the fusion head, then the wall-clock per-sample cost of the modality stage reduces from $\tilde{O}(\sum_i d_i)$ to $\tilde{O}(\max_i d_i)$. The (typically cheap) fusion pass adds $\tilde{O}(K)$, which is dominated when $K \le \max_i d_i$ or absorbed in the $\tilde{O}(\cdot)$ notation. Hence the wall-clock per-sample time can drop from $\tilde{O}(D)$ to $\tilde{O}(\max_i d_i)$, as claimed.

□

Proof. (Proof of Lemma 2) Let $\phi(x) := \frac{dQ_X}{dP_X}(x)$. The Lagrangian is $\mathcal{L} = \mathbb{E}_{P_{Y|X}\times P_X}\big[\ell(f\circ g(X),Y)\phi(X)\big] - \lambda\big(\mathbb{E}_{P_X}[(\phi(X)-1)^2] - \mathcal{B}\big) - \eta(\mathbb{E}_{P_X}[\phi(X)] - 1)$ .. Maximizing pointwise over ϕ (see, e.g., Rockafellar and Wets [2009]) yields $\phi^*(X) = 1 + \frac{1}{2\lambda}\big(\ell(f\circ g(X),Y) - \eta\big)$. Optimizing over $\lambda \ge 0$ and η gives the closed form for χ^2-DRO: $r(f\circ g, P_X) = \mathbb{E}[\ell] + \sqrt{\mathcal{B}}\ \sqrt{\mathrm{Var}(\ell)}$. □

Proof. (Proof of Lemma 3) The derivative df/dX_i is $\left|\frac{df}{dX_i}\right| = \frac{|X_i - \bar{X}|}{\sqrt{n\sum_{i=1}^n X_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^n X_i\right)^2}} = \frac{|X_i-\bar{X}|}{\sqrt{n\sum_{i=1}^n (X_i-\bar{X})^2}}$, and $\|\nabla f\|_2 = \sum_{i=1}^n \frac{(X_i-\bar{X})^2}{\sqrt{n\sum_{i=1}^n (X_i-\bar{X})^2}} = \frac{1}{\sqrt{n}}$ □

Proof. (Proof of Lemma 4) Let $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$, expand S_n around $\mu = \mathbb{E}S_n$ by Taylor's theorem $\sqrt{S_n} = \sqrt{\mu} + \frac{S_n-\mu}{2\sqrt{\mu}} - \frac{\mathbb{E}(S_n-\mu)^2}{2\mu^{\frac{3}{2}}} + O((S_n-\mu)^3)$. It is known that $Var(S_n) = \frac{1}{n}\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)$ where $\mu_4 = \mathbb{E}[X_i - \mathbb{E}X_i]^4$ $\mathbb{E}\sqrt{S_n} = \sqrt{\mu} - \frac{\mathbb{E}(S_n-\mu)^2}{2\mu^{\frac{3}{2}}} + O(\mathbb{E}(S_n-\mu)^3)$, and $|\mathbb{E}\sqrt{S_n} - \sqrt{\mu}| \le \frac{\mathbb{E}(S_n-\mu)^2}{2\mu^{\frac{3}{2}}} + O(\mathbb{E}(S_n-\mu)^3) = \frac{1}{n}\frac{\left(\mu_4 - \frac{n-3}{n-1}\mu_2^2\right)}{2\mu^{\frac{3}{2}}} + O(\mathbb{E}(S_n-\mu)^3) \le \frac{C}{n}$. □

Proof. (Proof of Theorem 1) By the χ^2-duality derivation in the main text, $r(f\circ g, P_X) = \mathbb{E}_{P_{Y|X}\times P_X}[\ell(f\circ g(X),Y)] + \sqrt{\mathcal{B}}\ \ h(P_{Y|X}\times P_X)$, where $h(P_{Y|X}\times P_X) := \sqrt{\mathrm{Var}_{P_{Y|X}\times P_X}\big(\ell(f\circ g(X),Y)\big)}$. The same representation holds with P_X replaced by $\hat{P}_X$. Hence

$$\begin{aligned}
&\big|r(f\circ g, \hat{P}_X) - \mathbb{E}_{P_{Y|X}\times P_X} r(f\circ g, \hat{P}_X)\big| \\
&\quad \le \sqrt{\mathcal{B}}\,\big|h(P_{Y|X}\times \hat{P}_X) - \mathbb{E}_{P_{Y|X}\times P_X} h(P_{Y|X}\times \hat{P}_X)\big| \\
&\qquad + \Big|\frac{1}{n}\sum_{i=1}^n \ell_i - \mathbb{E}_{P_{Y|X}\times P_X}\big[\frac{1}{n}\sum_{i=1}^n \ell_i\big]\Big|,
\end{aligned}$$

where $\ell_i := \ell(f\circ g(X_i),Y_i)$ and we use that the outer expectation is over the i.i.d. draw of $\{(X_i,Y_i)\}_{i=1}^n$ from $P_{Y|X}\times P_X$.

- **Step 1: Concentration of the empirical standard deviation.** Let $F(\ell_1,\dots,\ell_n) = \Big(\frac{1}{n}\sum_{i=1}^n(\ell_i - \bar{\ell})^2\Big)^{1/2}$ with $\bar{\ell} = \frac{1}{n}\sum_i \ell_i$. By Lemma 2 (Lipschitz property of empirical standard deviation), F is $n^{-1/2}$-Lipschitz w.r.t. the ℓ_2 norm. If $\ell(\cdot) \in [0, M_\ell]$, then by the concentration inequality in Lemma 1 (Boucheron–Lugosi–Massart), for all $t > 0$, with probability at least $1 - 2e^{-t}$,
$$\big|h(P_{Y|X}\times \hat{P}_X) - \mathbb{E}_{P_{Y|X}\times P_X} h(P_{Y|X}\times \hat{P}_X)\big| \ \le\ M_\ell\sqrt{\tfrac{2t}{n}}.$$

- **Step 2: Concentration of the empirical mean.** Since $\ell(\cdot) \in [0, M_\ell]$, Hoeffding's inequality yields, with probability at least $1 - 2e^{-t}$,

$$\Big|\tfrac{1}{n}\sum_{i=1}^{n} \ell_i - \mathbb{E}_{P_{Y|X}\times P_X}[\ell(f\circ g(X), Y)]\Big| \;\le\; M_\ell\sqrt{\tfrac{t}{2n}}.$$

- **Step 3: Combine.** Union bound gives the first displayed inequality in Theorem 1:

$$\big|r(f\circ g, \hat{P}_X) - \mathbb{E}_{P_{Y|X}\times P_X} r(f\circ g, \hat{P}_X)\big| \;\le\; \sqrt{\mathcal{B}}\, M_\ell\sqrt{\tfrac{2t}{n}} \;+\; M_\ell\sqrt{\tfrac{t}{2n}}.$$

- **Step 4: Bias between population and expected empirical risks.** Write

$$\begin{aligned}
&\big|\mathbb{E}_{P_{Y|X}\times P_X} r(f\circ g, \hat{P}_X) - r(f\circ g, P_X)\big| \\
&\quad= \sqrt{\mathcal{B}}\,\Big|\mathbb{E}_{P_{Y|X}\times P_X} h(P_{Y|X}\times\hat{P}_X) - h(P_{Y|X}\times P_X)\Big| \\
&\qquad+ \Big|\mathbb{E}_{P_{Y|X}\times P_X}\big[\tfrac{1}{n}\sum_{i=1}^{n}\ell_i\big] - \mathbb{E}_{P_{Y|X}\times P_X}[\ell(f\circ g(X), Y)]\Big|.
\end{aligned}$$

The mean term vanishes by linearity. By Lemma 3 (delta-method/Taylor expansion of $\sqrt{\cdot}$ around the population variance), there exists a universal constant C such that

$$\Big|\mathbb{E}\sqrt{\tfrac{1}{n}\sum_{i=1}^{n}(\ell_i - \bar{\ell})^2} - \sqrt{\mathrm{Var}(\ell)}\Big| \;\le\; \frac{C}{n}.$$

Multiplying by $\sqrt{\mathcal{B}}$ yields the second displayed inequality:

$$\big|\mathbb{E}_{P_{Y|X}\times P_X} r(f\circ g, \hat{P}_X) - r(f\circ g, P_X)\big| \;\le\; \sqrt{\mathcal{B}}\,\frac{C}{n}.$$

Finally, recall $\mathcal{B} = \sum_{k=1}^{K}\rho_k + 2\sum_{i<j}|\gamma_{ij}|\sqrt{\rho_i\rho_j}$ from the correlation-aware χ^2 bound, which gives the stated dependence.

□

Proof. (Proof Sketch of Theorem 2) Start from the triangle inequality to compare $r(g\circ\hat{f}, \widehat{P}_X)$ and $r(g\circ f, P_X)$ by introducing and subtracting $r(g\circ f, \widehat{P}_X)$, producing three natural gaps (encoder change at the same base, sampling fluctuation, and the population shift). For the encoder change, maximize over the same χ^2–ball around $\widehat{P}_X$ and apply the inequality $|\sup_Q F(Q) - \sup_Q G(Q)| \le \sup_Q |F(Q) - G(Q)|$ with $F(Q) = \mathbb{E}_Q[\ell(g\circ\hat{f}(X), Y)]$ and $G(Q) = \mathbb{E}_Q[\ell(g\circ f(X), Y)]$. The loss and head being Lipschitz give $|\ell(g\circ\hat{f}(X), Y) - \ell(g\circ f(X), Y)| \le L_\ell L_{g,i}|\Delta_i(X^{(i)})|$. A standard χ^2 change-of-measure bound then yields $\sup_{Q:\chi^2(Q\|\widehat{P}_X)\le B}\mathbb{E}_Q|\Delta_i(X^{(i)})| \;\le\; \mathbb{E}_{\widehat{P}_X}|\Delta_i(X^{(i)})| + \sqrt{\mathcal{B}\,\mathrm{Var}_{\widehat{P}_X}(|\Delta_i(X^{(i)})|)}$, giving the encoder term. □

Proof. (Proof of Theorem 3) We apply Le Cam's two-point method. Consider losses supported on $\{0, L\}$:

$$Z_1 = \begin{cases} 0 & \text{w.p. } 1 - p^* - \delta, \\ L & \text{w.p. } p^* + \delta, \end{cases} \qquad Z_2 = \begin{cases} 0 & \text{w.p. } 1 - p^* + \delta, \\ L & \text{w.p. } p^* - \delta, \end{cases}$$

with $p^* = \frac{1}{2}$ and $\delta \in (0, \frac{1}{2})$. For either distribution, $r(f\circ g, P_X) = \mu(\ell) + \sqrt{\mathcal{B}}\;\sigma(\ell), \quad \mu(\ell) = \mathbb{E}[Z], \;\; \sigma^2(\ell) = \mathrm{Var}(Z) = p(1-p)L^2$. At $p^* = \frac{1}{2}$, $\sigma(\ell) = \frac{L}{2}$ for both Z_1 and Z_2, so the variance term cancels and $|r(f\circ g, P_{X,1}) - r(f\circ g, P_{X,2})| = |\mu_1 - \mu_2| = 2L\delta =: s$. Le Cam's inequality then gives, for any estimator based on n samples, $\inf_{\widehat{r}}\ \sup_{P\in\{P_{X,1}, P_{X,2}\}}\mathbb{E}\big[\,|\widehat{r} - r(f\circ g, P)|\,\big] \;\ge\; \frac{s}{2}\Big(1 - \|P^n_{X,1} - P^n_{X,2}\|_{\mathrm{TV}}\Big)$. By Pinsker's inequality, $\|P^n_{X,1} - P^n_{X,2}\|_{\mathrm{TV}} \;\le\; \sqrt{\frac{n}{2}\,D_{\mathrm{kl}}(P_{X,2}\|P_{X,1})}$, and for the Bernoulli pair above (with parameters $\frac{1}{2} \pm \delta$),

$$D_{\mathrm{kl}}(P_{X,2}\|P_{X,1}) = \Big(\tfrac{1}{2} + \delta\Big)\log\frac{\frac{1}{2} + \delta}{\frac{1}{2} - \delta} + \Big(\tfrac{1}{2} - \delta\Big)\log\frac{\frac{1}{2} - \delta}{\frac{1}{2} + \delta} = 2\delta\log\frac{\frac{1}{2} + \delta}{\frac{1}{2} - \delta}.$$

Choose $\delta = \frac{1}{2n \log(M-1)}$ for a fixed $M > 2$. Then $D_{\mathrm{kl}}(P_{X,2} \| P_{X,1}) \leq \frac{1}{n}$ (by $\log \frac{\frac{1}{2}+\delta}{\frac{1}{2}-\delta} \leq \log(M-1)$ under the stated choice), and hence $\|P_{X,1}^n - P_{X,2}^n\|_{\mathrm{TV}} \leq \frac{1}{2}$ for sufficiently large n. With $s = 2L\delta = \frac{L}{n \log(M-1)}$ we conclude

$$\inf_{\widehat{r}} \sup_{P \in \mathcal{P}} \mathbb{E}\big[\,|\widehat{r} - r(f \circ g, P)|\,\big] \;\geq\; \frac{s}{2}\Big(1 - \tfrac{1}{2}\Big) = \frac{L}{4n \log(M-1)}.$$

This establishes the stated lower bound. □

C Additional Computational Results

C.1 Dataset Descriptions

Healthcare The healthcare dataset HAIM-MIMIC-MM is a multimodal dataset of 34,537 samples that contains 7279 unique hospitalizations and 6485 patients. It contains four modalities: tabular, time-series, text and images. We consider two binary predictive tasks: mortality in the next 48 hours, and discharge in the next 48 hours. Specifically, embeddings were generated for each of these modalities, with details to be found in Soenksen et al. [2022].

Language. The Huffpost dataset contains approximately 200,000 samples of news article headlines from 11 categories (Black Voices, Business, Comedy, Crime, Entertainment, Impact, Queer Voices, Science, Sports, Tech, Travel). The target is to identify these category tags from the original headline text. The dataset dates from 2012 to 2018, and samples accumulated per year is considered a single individual time period. We consider three modality-specific blocks: headline, short-description embeddings, and metadata. We process each sample from the original language format to a respective embedding using the BAAI FlagEmbedding model to lower-dimensional vector of size 2014. We then use this embedding for the downstream prediction problem.

C.2 χ^2-divergence and Correlation

We consider two Gaussian distributions with the *same* covariance and a mean shift: $P = \mathcal{N}(\mu_P, \Sigma)$ and $Q = \mathcal{N}(\mu_Q, \Sigma)$ with $\mu_Q = \mu_P + \Delta$, $\mu_P = \mathbf{0}$, and $\Sigma = \begin{pmatrix} \sigma_1^2 & c\,\sigma_1\sigma_2 \\ c\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}, \quad |c| < 1$. We fix $\sigma_1 = \sigma_2 = 1$ and draw a standardized mean shift $z \sim \mathcal{N}(0, 0.5^2 I_2)$ independently each trial, then set $\Delta = z \odot (\sigma_1, \sigma_2)$. For each correlation $c \in \{-0.6, -0.3, 0, 0.3, 0.6\}$ we run 200 trials.

c	mean	std	min	max
-0.6	5.288	29.283	0.005	379.781
-0.3	1.286	3.919	0.004	50.341
0.0	0.875	2.230	0.000	25.456
0.3	1.229	2.394	0.001	22.107
0.6	4.499	16.556	0.009	166.085

Table 3: χ^2-divergence statistics over 200 trials with $\sigma_1 = \sigma_2 = 1$ and $z \sim \mathcal{N}(0, 0.5^2 I_2)$.

In multimodal learning, stronger cross-modal correlation—whether positive or negative—magnifies how covariate shifts combine across modalities, leading to a larger chi-square divergence between source and target. When the modalities are uncorrelated, the divergence is smallest.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and introduction both explicitly list the contributions of the paper (lines 4-11 and section 1 paragraph 3). Theoretical claims (i.e. justification of modality-aware formulation, generalization bound and risk lower bounds etc.) are supported via theorems and proofs, and conclusions about empirical performance are supported by our simulations and 3 real world case studies.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: The limitations of the paper are discussed throughout the paper (e.g., consideration of other distance metric such as Wasserstein) and summarized in the Conclusion.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: Assumptions for all theoretical results are stated explicitly in the theorem, lemma, or proposition statements, and each theorem or lemma is immediately followed by a proof (or a reference to one).
4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: The contribution of the paper is mainly the new DRO framework along with its theoretical results. Experimental details are also given in Section 4 as well as the Appendix C.
5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: The vision and language datasets are obtained from the public data base WildTime, which is availale to the public from their repository `https://wild-time.github.io/`. The healthcare data is obtained from MIMIC. The code base will be released/open-sourced on Github upon completion of the review of the paper.
6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?
   Answer: [Yes]
   Justification: All descriptions of the computational experiments are clearly articulated in Section 4 and Appendix C.
7. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: We provide detailed computational resource statement at the beginning of Section 4.
8. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: The research conforms to each of the guidelines listed in the NeurIPS Code of Ethics.
9. **Broader impacts**
   Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
   Answer: [Yes]

Justification: The potential applications and societal impact of the proposed methodology are discussed in the introduction. The theoretical guarantees and their implications to high-stake settings such as healthcare are also remarked in section 3.

10. **Safeguards**
    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?
    Answer: [NA]
    Justification: The contributions of the work do not include release of data or models.
11. **Licenses for existing assets**
    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
    Answer: [NA]
    Justification: The paper does not use existing assets.
12. **New assets**
    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
    Answer: [NA]
    Justification: The contributions of the paper do not include the release of new assets.
13. **Crowdsourcing and research with human subjects**
    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
    Answer: [NA]
    Justification: This paper does not involve crowdsourcing or research with human subjects.
14. **Institutional review board (IRB) approvals or equivalent for research with human subjects**
    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
    Answer: [NA]
    Justification: The study did not use data requiring IRB approvals.
15. **Declaration of LLM usage**
    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.
    Answer:
    answerNA
    Justification: The paper's usage of LLM is only through the processing of embedding in the language dataset and does not constitute a core component of the methodology.