ShapKO: Shapley-Adaptive Modality Knockout for Robust Multimodal Learning

Nusrat Binta Nizam[1,2,3], Fengbei Liu[1,3], Sunwoo Kwak[1,2,3], Minh Nguyen[1,2,3], Ruining Deng[3], and Mert R. Sabuncu[1,2,3]

[1] Cornell University, New York, NY, USA
[2] Cornell Tech, New York, NY, USA
[3] Weill Cornell Medicine, New York, NY, USA
{nn284,fl453,sk3355,bn244,msabuncu}@cornell.edu, rud4004@med.cornell.edu

Abstract. Multimodal medical models often degrade when inputs are missing, a common scenario in real-world clinical workflows. Separately, even when all modalities are present, modality dominance is observed during training, where optimization over-relies on a highly predictive modality and undertrains complementary sources, resulting in poor robustness under partial availability. While training-time modality knockout improves missing-modality robustness, existing approaches use static masking rates that cannot adapt to evolving modality utility during training. We introduce ShapKO (Shapley-Adaptive Modality Knockout), a dynamic training strategy that learns modality-specific knockout probabilities based on validation utility. ShapKO periodically evaluates performance across modality subsets, estimates modality importance via Shapley values, and updates masking probabilities to suppress dominant modalities more frequently. This adaptive process promotes complementary representations, while requiring no architectural modifications. We evaluate ShapKO on three datasets covering multitask clinical classification, survival prediction, and cancer detection. ShapKO consistently improves performance under modality absence and yields interpretable trajectories of learned masking behavior. Code is available at: https://github.com/sumona00/ShapKO.

Keywords: Robust Multimodal learning · missing modalities · Shapley value · Knockout

1 Introduction

Multimodal learning aims to improve clinical decision support by combining complementary evidence from heterogeneous sources such as imaging, physiological signals, laboratory measurements, and structured clinical variables. In practice, modalities are often missing due to cost and triage, conditional ordering, acquisition failures, and site-specific protocols [29]. Consequently, models trained under the assumption that all modalities are present at inference can degrade substantially under realistic missingness [18].

A key contributor is *modality dominance*: when one modality is consistently most predictive during training, optimization over-relies on it while undertraining weaker but complementary modalities, so that the model collapses onto the dominant modality and fails when it is unavailable at inference [7,31]. Addressing this imbalance requires maintaining performance across diverse modality-availability patterns without training separate subset-specific models [12]. Existing robustness strategies, including modality masking, knockout, and gradient balancing, attempt to mitigate this issue by simulating missingness or reweighting modality contributions during training [19,29,20,17]. However, these approaches typically rely on static heuristics or local training signals and do not explicitly optimize performance across the combinatorial space of modality subsets. As a result, they cannot adapt to evolving modality utility dynamics and may over-regularize weaker modalities or insufficiently suppress dominant ones.

Missing modalities are common in clinical data and can substantially degrade multimodal performance, motivating methods that operate under arbitrary availability patterns [29]. In medical imaging, prior work includes any-subset fusion in a shared latent space for inference from partial modality sets [6] and generative/shared-latent formulations that support prediction and modality completion [2,28]. Common baselines impute missing inputs or train subset-specific models, but imputation can introduce distribution shift and subset models scale poorly with modality count [3]. A widely used, architecture-agnostic alternative is training-time modality masking/knockout [19,18,13], yet static masking rates can exacerbate modality dominance and related collapse phenomena linked to optimization effects such as gradient conflict [7]. Beyond masking, modality-balancing approaches (e.g., OGM-GE [21]) and configuration-aware models improve robustness across modality subsets [21,20,5,10].

We propose ShapKO, which learns per-modality knockout probabilities from a validation *utility* evaluated on modality subsets. While standard signals such as validation loss or accuracy on the *all-modality* input provide a single-configuration view that can be dominated by the strongest modality, our goal is robustness across the many availability patterns encountered at deployment [33,8]. We therefore estimate each modality's average marginal contribution across subsets using Shapley values, a principled coalition-based attribution that accounts for redundancy and complementarity between modalities [25]. ShapKO periodically evaluates utilities over subsets, estimates modality importance, and updates knockout probabilities with a stable *knockout-strong-more* rule that masks more influential modalities more often, reducing modality dominance and improving robustness under missingness. Our contributions are as follows:

- We propose ShapKO, a utility-driven, architecture-agnostic modality *knockout* method that adapts per-modality knockout probabilities during training based on Shapley-estimated marginal contributions. To our knowledge, ShapKO is the first method that integrates Shapley-values into training for optimization.
- We evaluate ShapKO in three settings - multitask clinical classification, multimodal prostate cancer detection, and survival prediction - showing im-

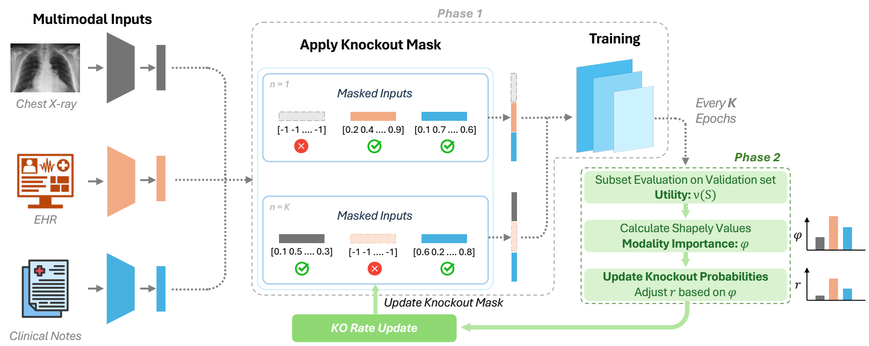

Fig. 1. Overview of ShapKO (architecture-agnostic). ShapKO trains with sampled modality masks by replacing knocked-out modalities (or embeddings) with fixed placeholders before fusion. Every K epochs, it evaluates validation utility over modality subsets, estimates Shapley-based modality importance, and updates per-modality knockout probabilities with a stable *drop-strong-more* rule. The procedure is architecture-agnostic and applies across tasks.

proved robustness compared to standard training, fixed-rate knockout (Fixed KO) [19], OGM-GE, ModDrop++ [14], PMR [4], and SimMLM [11], providing interpretable knockout-rate trajectories over training.

2 Methodology

Problem setup. Let $\mathcal{M} = \{1, \ldots, d\}$ be the full modality index set and $\mathcal{O} \subseteq \mathcal{M}$ the modalities available in the dataset. Given a training sample (x^i, y^i), each observed modality $m \in \mathcal{O}^i \subseteq \mathcal{O}$ is encoded by a modality-specific encoder g_m into an embedding $e^i_m = g_m(x^i_m) \in \mathbb{R}$, giving $e^i = \{e^i_m\}_{m \in \mathcal{O}^i}$. ShapKO aims to learn a predictor f_θ that takes the *set* of modality embeddings (with placeholders substituted for absent modalities), fuses them, and outputs a single task prediction $\hat{y}^i = f_\theta(\{e^i_m\}_{m \in \mathcal{M}})$, and that is robust under arbitrary modality availability during inference. To achieve this, ShapKO introduces additional synthetic knockout during training: a random subset $\mathcal{A}^i \subseteq \mathcal{O}^i$ of present modalities is further masked independently of (x^i, y^i).

Modalities in $\mathcal{M} \setminus \mathcal{O}^i$ are *structurally missing* due to heterogeneous clinical acquisition, and modalities in $\mathcal{O}^i \setminus \mathcal{A}^i$ are *synthetically absent* due to training-time knockout. Both missingness are handled by replacing the corresponding embeddings with different fixed placeholders $\tilde{e}_m$, where the choice of $\tilde{e}_m$ is given in Sec. 3.

2.1 Utility Estimation and Knockout Rate Update

Utility function. We treat modalities in $\mathcal{M}$ as players in a cooperative game [25]. For any subset $S \subseteq \mathcal{O}$, we define a scalar utility $v(S)$ by evaluating the current model on the validation set using only modalities in S, with all $m \notin S$ replaced by $\tilde{e}_m$. Let $J(S)$ denote the task-specific evaluation metric. We define $v(S) = J(S)$ for higher-is-better metrics (e.g., AUC, Accuracy, C-index) and $v(S) = -J(S)$ for lower-is-better metrics (e.g., validation loss), so that larger $v(S)$ always indicates better performance regardless of metric direction. The Shapley value of modality m is:

$$\phi_m = \sum_{S \subseteq \mathcal{O} \setminus \{m\}} \frac{|S|!\,(|\mathcal{O}| - |S| - 1)!}{|\mathcal{O}|!} \Big(v(S \cup \{m\}) - v(S) \Big), \tag{1}$$

where ϕ_m is the weighted average marginal gain from adding m across all coalitions [25], capturing both its individual contribution and interactions with other modalities. A larger ϕ_m indicates that modality m is more dominant. We instantiate $J(\cdot)$ per dataset in Sec. 3.

Shapley-Adaptive Knockout Rates. We convert Shapley values to weights on the probability simplex by shifting to non-negative values and normalizing:

$$w_m = \frac{\phi_m - \phi_{\min}}{\sum_{m' \in \mathcal{O}} \phi_{m'} - \phi_{\min}}, \quad \phi_{\min} = \min_{m \in \mathcal{O}} \phi_m, \tag{2}$$

so that $w_m \geq 0$ and $\sum_{m \in \mathcal{O}} w_m = 1$. Let $\bar{w} = 1/|\mathcal{O}|$ be the uniform weight. Per-modality knockout probabilities are then updated as:

$$\begin{gathered} r_m = \operatorname{clip}\left(r \left(\frac{w_m}{\bar{w}} \right)^{\delta}, r_{\min}, r_{\max} \right), \\ (1 - r)^d = 0.5 \;\Rightarrow\; r = 1 - 0.5^{1/d}, \end{gathered} \tag{3}$$

where $\delta \geq 0$ controls sensitivity (lower for higher structural missingness). r is the base knockout rate initialized to random based on number of modalities. $[r_{\min}, r_{\max}]$ bounds the rates for stability. Modalities with $w_m > \bar{w}$ (above-average importance) receive a higher knockout rate than the base r, directly implementing the *drop-strong-more* policy.

2.2 Training Description

ShapKO alternates between two phases: optimizing f_θ under the current knockout distribution (**Phase 1**), and updating the knockout rates based on Shapley-estimated modality importance (**Phase 2**).

In **Phase 1**, for each modality $m \in \mathcal{O}_i$ we sample $z_m \sim \text{Bernoulli}(1 - r_m)$ (with $\sum_m z_m \geq 1$ enforced) and form the knocked-out embedding $\hat{e}^i_m$,

$$\hat{e}^i_m = \begin{cases} z_m\, e^i_m + (1 - z_m)\, \tilde{e}_m, & m \in \mathcal{O}_i, \\ \tilde{e}_m, & m \in \mathcal{M} \setminus \mathcal{O}_i, \end{cases} \tag{4}$$

Table 1. Hyperparameters and placeholder values used across tasks.

	Epochs	Learning rate	Weight decay	Optimizer	K	d	$r_{\min}$	$r_{\max}$	δ	Struct. placeholder	Synth. placeholder
Prostate MRI	100	3e-4	1e-4	AdamW	15	3	0.02	0.30	0.5	0	0
Survival Prediction	300	2e-4	5e-5	AdamW	20	4	0.02	0.60	0.5	0	-1
Multitask Classif.	100	1e-4	1e-5	AdamW	10	3	0.02	0.35	1.0	`[MISS]`	`[KO]`

Collecting the knocked-out embeddings of all modalities into the set $\hat{e}^i = \{\hat{e}^i_m\}_{m\in\mathcal{M}}$, the network f_θ fuses $\hat{e}^i$ into a single prediction, and θ is trained by minimizing the empirical task loss over the N training samples:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\big(f_\theta(\hat{e}^i),\ y^i\big)\,. \tag{5}$$

In **Phase 2**, triggered every K epochs after warm-up, the current f_θ evaluates $v(S)$ over all subsets in validation split, estimates ϕ_m (Eqs. (1)), and updates r_m (Eqs. (2)–(3)), with θ fixed.

3 Experimental Design

3.1 Dataset and Implementation details

Dataset organization. We evaluate on three multimodal benchmarks covering complementary clinical tasks. **FlexCare** integrates structured EHR variables, chest imaging, and clinical text (MIMIC-IV/CXR/NOTE) for multi-task clinical classification [30]. For time-to-event modeling, we use the **MMD** benchmark combining radiology, pathology, genomics, and demographics for survival prediction [1]. For imaging robustness, we use **prostate MRI datasets** with T2w, DWI, ADC modalities for binary clinically significant prostate cancer detection [22].

Implementation details. For Prostate MRI datasets, we use U-Net style 3D CNNs as modality encoder g_m. Embeddings are concatenated and fed to a shallow fusion MLP for binary prediction, trained with binary cross-entropy loss [15]. For $J(\cdot)$ we use AUC as metric. We use a fixed zero placeholder $\tilde{e}_m$ for both structural and synthetic missingness. For MMD, we follow previous work [1] and use the same modality-specific encoders for pathology, radiology, genomics, and demographics. We use C-index for $J(\cdot)$, the standard metric for survival prediction. We adopt a fixed zero placeholder for structural missingness and -1 placeholder for synthetic missingness. We optimize the model using the Cox partial-likelihood loss as formulated in [1,23]. For FlexCare, we follow their setting and adopt the tri-modal Transformer architecture over EHR time-series, chest X-ray (CXR) images, and clinical notes [30]. We select averaged multi-task AUC as $J(\cdot)$. We use a learned `[MISS]` token embedding for structural missingness and a separate `[KO]` token embedding for synthetic missingness, with attention masking

Table 2. AUC (%, mean±std over 5 folds) for binary classification under different modality combinations. Best method per row is bolded.

Modality subset	Baseline	OGM-GE	Fixed KO	ShapKO
T2w	0.656±0.127	0.636 ± 0.031	**0.685±0.048**	0.663±0.067
ADC	0.732±0.084	0.726 ± 0.065	0.738±0.095	**0.739±0.143**
DWI	0.773±0.157	0.787 ± 0.041	0.818±0.055	**0.842±0.055**
T2w+ADC	0.747±0.082	0.724 ± 0.067	0.765±0.070	**0.775±0.113**
T2w+DWI	0.777±0.150	0.787 ± 0.040	0.824±0.045	**0.855±0.036**
ADC+DWI	0.831±0.025	0.819 ± 0.018	0.839±0.040	**0.878±0.038**
T2w+ADC+DWI	0.831±0.026	0.818 ± 0.017	0.840±0.038	**0.887±0.029**

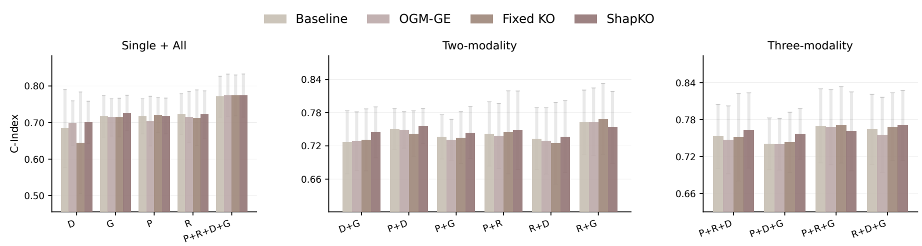

Fig. 2. C-index across modality subsets for Baseline, OGM-GE, Fixed KO, and ShapKO (15 CV folds); gray ranges show ±std. Modalities: D = Demographics, G = Genomics, P = Pathology, R = Radiology; combinations denote fusion (e.g., P+R+D+G uses all four).

to prevent information leakage under partial availability [24]. Note we drop the Mixture of Experts (MoE) component in the original FlexCare implementation for fair comparison with other baselines. We train this model with a task-specific supervised loss (binary cross-entropy for most tasks and cross-entropy for multi-class tasks such as length-of-stay) [15].

Hyperparameters. Table 1 provides the hyperparameters and placeholder values for different tasks. All hyperparameters were set empirically and refined by cross-validation on the training folds, selecting values that yielded stable learning dynamics and the best validation performance.

4 Results and Discussions

Across all benchmarks, modality knockout improves robustness to missing inputs by reducing train–test mismatch and mitigating modality dominance. In particular, ShapKO consistently yields the strongest performance under partial modality availability, indicating that utility-driven knockout can better calibrate reliance on modalities as training progresses.

Prostate Cancer Detection. Table 2 summarizes binary classification AUC under all modality combinations of T2w, ADC, and DWI. Fixed KO increases AUC over the baseline and OGM-GE for several subsets, while ShapKO achieves the best performance for most combinations. In particular, ShapKO achieves the highest AUC for all the combinations except T2w. Additionally, ShapKO achieves the highest AUC on 6 of 7 modality subsets, with a full-modality AUC of 0.887 against 0.814 (ModDrop++), 0.821 (PMR), and 0.811 (SimMLM). These results indicate that ShapKO strengthens performance when all modalities are available and yields consistent gains when deployed inputs are incomplete, a frequent scenario in prostate MRI due to missing sequences.

Survival Prediction. Figure 2 compares C-index values across all modality subsets (D,G,P,R). Fixed KO improves over the baseline in many missing-modality settings, and OGM-GE provides less consistent gains. ShapKO achieves performance that is consistently comparable to or better than Fixed KO across most modality subsets. Also, ShapKO is on par with PMR and ahead of ModDrop++ and SimMLM, with a mean C-index across 15 modality subsets of 0.741 (ShapKO), 0.740 (ModDrop++), 0.742 (PMR), and 0.738 (SimMLM). While the absolute improvements in C-index are modest, ShapKO avoids degradation in any configuration and provides more stable performance when prediction relies on less informative modality combinations. Importantly, ShapKO maintains baseline level performance in the full-modality setting while improving robustness under partial-modality evaluation, indicating that this training approach does not compromise overall survival modeling capacity. Figure 3 shows non-uniform ShapKO knockout-rate trajectories (mean±std), indicating adaptation to modality utility. The knockout-rate trajectories provide a simple diagnostic of modality utility drift over training, revealing non-uniform knockout behavior across modalities.

Multitask Classification. Figure 4 reports AUC for six clinical prediction tasks under modality subsets of EHR (E), chest X-ray (C), and clinical notes (N). Both Fixed KO and ShapKO improve over the baseline and OGM-GE in missing-modality settings, showing that training with modality knockout yields models that generalize better to partial inputs. ShapKO is consistently strongest across subsets, with especially clear gains in single-modality inference (E-only, C-only, or N-only), where robustness depends on how well each unimodal encoder is trained despite multimodal optimization. ShapKO is the strongest on 4 of 6 tasks under the full-modality setting, including in-hospital mortality (0.871 vs. 0.856/0.844/0.858 for ModDrop++/PMR/SimMLM), phenotyping (0.826 vs. 0.817/0.800/0.806), length-of-stay (0.777 vs. 0.760/0.752/0.753), and diagnosis (0.689 vs. 0.656/0.634/0.667). These results are consistent with ShapKO mitigating modality dominance by adaptively increasing knockout on influential modalities over training, encouraging better use of complementary signals when other modalities are absent.

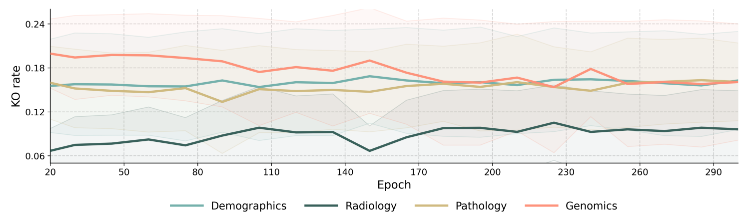

Fig. 3. Learned ShapKO knockout rates over training (K=20), shown per modality as mean (line) ±std (band) across 15 CV folds at KO-update epochs. Higher rate = more frequent knockout.

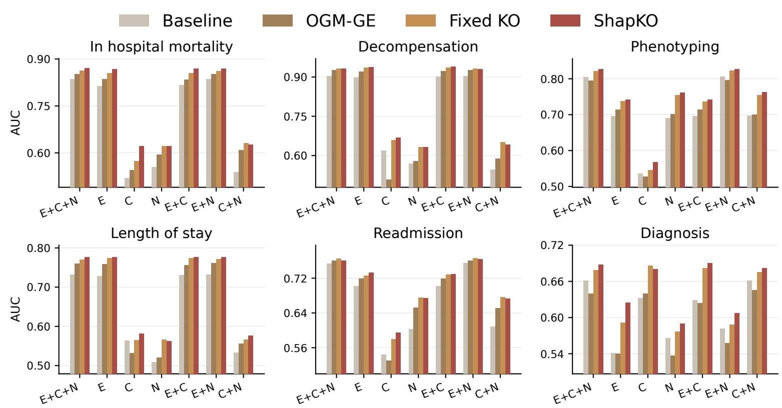

Fig. 4. AUC on MIMIC-IV multitask classification across modality subsets for Baseline, OGM-GE, Fixed KO, and ShapKO. Tasks: in-hospital mortality, decompensation, phenotyping, length of stay, readmission, diagnosis. Modalities: E = EHR, C = chest X-ray, N = clinical notes (E+C+N uses all three).

5 Limitations and Future Work

As future work, the validation-time overhead of ShapKO could be reduced by replacing exhaustive subset evaluation (including task-specific metrics) with approximate Shapley estimation, e.g., subset sampling, permutation-based Monte Carlo, or surrogate explainers such as KernelSHAP [26,16]; out-of-bag utility estimation [9] could further remove the dependence on a separate validation split, while preserving accurate modality-importance tracking. ShapKO also depends on design choices such as the structural vs. synthetic placeholder split, distinguishing clinically absent modalities from artificially knocked-out ones, which may not be universally optimal — the subset used for utility estimation, and hyperparameters, which may affect stability and require tuning across datasets. Additionally, ShapKO optimizes average subset utility and may not explicitly prioritize worst-case or clinically prevalent missingness patterns, like, Flex-MoE [32], REMIND [27]; future work could incorporate distribution or risk-aware update rules and replace fixed placeholders with trainable missing/knockout embeddings learned end-to-end. Also, ShapKO targets missing modalities; robustness to corrupted-but-present inputs requires explicit corruption detection and is left to future work. Finally, the utility and knockout-update mechanism are head-agnostic and apply to voxel-level segmentation via a Dice-based J(S), making dense-prediction benchmarks a natural extension.

6 Conclusion

We presented ShapKO, a Shapley-guided modality knockout strategy that dynamically adjusts knockout rates to reduce modality dominance and improve robustness to missing inputs. Across multitask classification, survival prediction, and prostate cancer detection, ShapKO consistently improved performance under partial-modality settings and often matched or exceeded full-modality baselines. These results highlight the value of utility-driven, training-time modality knockout for reliable multimodal learning in real-world clinical deployments. Moreover, ShapKO yields interpretable knockout-rate trajectories that provide a direct diagnostic of evolving modality utility and reliance during training.

Acknowledgments. This work is supported by NYP–Cornell Cardiovascular AI Collaboration. Additionally, this work is partially supported by National Institute of Health (NIH) awards: 1R01HL174863-01A1, and 1U54DK144866-01.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cui, C., Liu, H., Liu, Q., Deng, R., Asad, Z., Wang, Y., Zhao, S., Yang, H., Landman, B.A., Huo, Y.: Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 626–635. Springer (2022)
2. Dorent, R., Joutard, S., Modat, M., Ourselin, S., Vercauteren, T.: Hetero-modal variational encoder-decoder for joint modality completion and segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 74–82. Springer (2019)
3. Enders, C.K., Baraldi, A.N.: Missing data handling methods. The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development pp. 139–185 (2018)
4. Fan, Y., Xu, W., Wang, H., Wang, J., Guo, S.: Pmr: Prototypical modal rebalance for multimodal learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20029–20038 (2023)
5. Guo, Z., Jin, T., Chen, J., Zhao, Z.: Classifier-guided gradient modulation for enhanced multimodal learning. Advances in Neural Information Processing Systems **37**, 133328–133344 (2024)

6. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: Hemis: Hetero-modal image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 469–477. Springer (2016)
7. Javaloy, A., Meghdadi, M., Valera, I.: Mitigating modality collapse in multimodal vaes via impartial optimization. In: International Conference on Machine Learning. pp. 9938–9964. PMLR (2022)
8. Kontras, K., Chatzichristos, C., Blaschko, M., De Vos, M.: Improving multimodal learning with multi-loss gradient modulation. arXiv preprint arXiv:2405.07930 (2024)
9. Kwon, Y., Zou, J.: Data-oob: Out-of-bag estimate as a simple and efficient data value. In: International conference on machine learning. pp. 18135–18152. PMLR (2023)
10. Li, H., Li, X., Hu, P., Lei, Y., Li, C., Zhou, Y.: Boosting multi-modal model performance with adaptive gradient modulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22214–22224 (2023)
11. Li, S., Chen, C., Han, J.: 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)
12. Li, S., Du, C., Zhao, Y., Huang, Y., Zhao, H.: What makes for robust multi-modal models in the face of missing modalities? arXiv preprint arXiv:2310.06383 (2023)
13. Liu, G.H., Siravuru, A., Prabhakar, S., Veloso, M., Kantor, G.: Learning end-to-end multimodal sensor policies for autonomous navigation. In: Conference on robot learning. pp. 249–261. PMLR (2017)
14. Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H.H., Zhang, H., Oguz, I.: Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 444–453. Springer (2022)
15. Mannor, S., Peleg, D., Rubinstein, R.: The cross entropy method for classification. In: Proceedings of the 22nd international conference on Machine learning. pp. 561–568 (2005)
16. Mitchell, R., Cooper, J., Frank, E., Holmes, G.: Sampling permutations for shapley value estimation. Journal of Machine Learning Research **23**(43), 1–46 (2022)
17. Mohta, A., Chou, F.C., Becker, B.C., Vallespi-Gonzalez, C., Djuric, N.: Investigating the effect of sensor modalities in multi-sensor detection-prediction models. arXiv preprint arXiv:2101.03279 (2021)
18. Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1692–1706 (2015)
19. Nguyen, M., Karaman, B.K., Kim, H., Wang, A.Q., Liu, F., Sabuncu, M.R.: Knockout: A simple way to handle missing inputs. Transactions on machine learning research **2025**, 4468 (2025)
20. Novosad, P., Carano, R.A., Krishnan, A.P.: A task-conditional mixture-of-experts model for missing modality segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 34–43. Springer (2024)
21. Peng, X., Wei, Y., Deng, A., Wang, D., Hu, D.: Balanced multimodal learning via on-the-fly gradient modulation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8238–8247 (2022)
22. Saha, A., Bosma, J.S., Twilt, J.J., Van Ginneken, B., Bjartell, A., Padhani, A.R., Bonekamp, D., Villeirs, G., Salomon, G., Giannarini, G., et al.: Artificial intelligence and radiologists in prostate cancer detection on mri (pi-cai): an international,

paired, non-inferiority, confirmatory study. The Lancet Oncology **25**(7), 879–887 (2024)

23. Sasieni, P.: Maximum weighted partial likelihood estimators for the cox model. Journal of the American Statistical Association **88**(421), 144–152 (1993)
24. Shi, J., Yu, L., Cheng, Q., Yang, X., Cheng, K.T., Yan, Z.: Mftrans: Modality-masked fusion transformer for incomplete multi-modality brain tumor segmentation. IEEE journal of biomedical and health informatics **28**(1), 379–390 (2023)
25. Winter, E.: The shapley value. Handbook of game theory with economic applications **3**, 2025–2054 (2002)
26. Witter, R.T., Liu, Y., Musco, C.: Regression-adjusted monte carlo estimators for shapley values and probabilistic values. arXiv preprint arXiv:2506.11849 (2025)
27. Wu, C., Shuai, Z., Shen, L.: Remind: Rethinking medical high-modality learning under missingness–a long-tailed distribution perspective. arXiv preprint arXiv:2603.00046 (2026)
28. Wu, M., Goodman, N.: Multimodal generative models for scalable weakly-supervised learning. Advances in neural information processing systems **31** (2018)
29. Wu, R., Wang, H., Chen, H.T., Carneiro, G.: Deep multimodal learning with missing modality: A survey. arXiv preprint arXiv:2409.07825 (2024)
30. Xu, M., Zhu, Z., Li, Y., Zheng, S., Zhao, Y., He, K., Zhao, Y.: Flexcare: Leveraging cross-task synergy for flexible multimodal healthcare prediction. In: Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 3610–3620 (2024)
31. Yang, Y., Wan, F., Jiang, Q.Y., Xu, Y.: Facilitating multimodal classification via dynamically learning modality gap. Advances in Neural Information Processing Systems **37**, 62108–62122 (2024)
32. Yun, S., Choi, I., Peng, J., Wu, Y., Bao, J., Zhang, Q., Xin, J., Long, Q., Chen, T.: Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. Advances in Neural Information Processing Systems **37**, 98782–98805 (2024)
33. Zhang, X., Yoon, J., Bansal, M., Yao, H.: Multimodal representation learning by alternating unimodal adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 27456–27466 (2024)